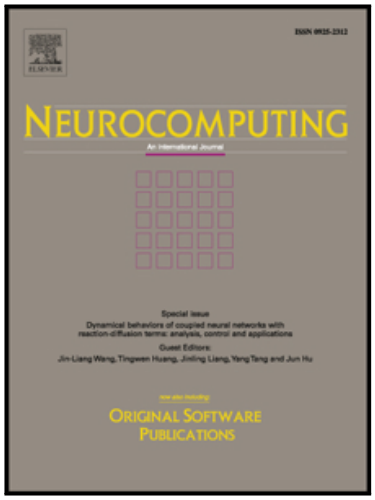
Communicated by Dr. Swagatam Das

Accepted Manuscript



Theoretical Foundations of Forward Feature Selection Methods based on Mutual Information

Francisco Macedo, M. Rosário Oliveira, António Pacheco, Rui Valadas

PII: S0925-2312(18)31158-5
DOI: https://doi.org/10.1016/j.neucom.2018.09.077
Reference: NEUCOM 20010

To appear in: *Neurocomputing*

Received date: 15 February 2018
Revised date: 21 June 2018
Accepted date: 28 September 2018

Please cite this article as: Francisco Macedo, M. Rosário Oliveira, António Pacheco, Rui Valadas, Theoretical Foundations of Forward Feature Selection Methods based on Mutual Information, *Neurocomputing* (2018), doi: https://doi.org/10.1016/j.neucom.2018.09.077

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Highlights

- Theoretical framework for the comparison of feature selection methods.
- Derivation of upper and lower bounds for the target objective functions.
- Linking objective function bounds with feature types.
- Distributional setting to highlight deficiencies of feature selection methods.
- Identification of feature selection methods to be avoided and preferred.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Theoretical Foundations of Forward Feature Selection Methods based on Mutual Information

Francisco Macedo[a,b], M. Rosário Oliveira[a,*], António Pacheco[a], Rui Valadas[c]

[a] *CEMAT and Department of Mathematics, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal*
[b] *EPF Lausanne, SB-MATHICSE-ANCHP, Station 8, CH-1015 Lausanne, Switzerland*
[c] *IT and Departament of Electrical and Computer Engineering, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal*

Abstract

Feature selection problems arise in a variety of applications, such as microarray analysis, clinical prediction, text categorization, image classification and face recognition, multi-label learning, and classification of internet traffic. Among the various classes of methods, forward feature selection methods based on mutual information have become very popular and are widely used in practice. However, comparative evaluations of these methods have been limited by being based on specific datasets and classifiers. In this paper, we develop a theoretical framework that allows evaluating the methods based on their theoretical properties. Our framework is grounded on the properties of the target objective function that the methods try to approximate, and on a novel categorization of features, according to their contribution to the explanation of the class; we derive upper and lower bounds for the target objective function and relate these bounds with the feature types. Then, we characterize the types of approximations taken by the methods, and analyze how these approximations cope with the good properties of the target objective function. Additionally, we develop a distributional setting designed to illustrate the various deficiencies of

*Corresponding author
Email addresses: francisco.macedo@epfl.ch (Francisco Macedo), rosario.oliveira@tecnico.ulisboa.pt (M. Rosário Oliveira), apacheco@math.tecnico.ulisboa.pt (António Pacheco), rui.valadas@tecnico.ulisboa.pt (Rui Valadas)

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

the methods, and provide several examples of wrong feature selections. Based on our work, we identify clearly the methods that should be avoided, and the methods that currently have the best performance.

Keywords:
Mutual information, Feature selection methods, Forward greedy search, Performance measure, Minimum Bayes risk

1. Introduction

In an era of data abundance, of a complex nature, it is of utmost importance to extract from the data useful and valuable knowledge for real problem solving. Companies seek in the pool of available information commercial value that can leverage them among competitors or give support for making strategic decisions. One important step in this process is the selection of relevant and non-redundant information in order to clearly define the problem at hand and aim for its solution [1].

Feature selection problems arise in a variety of applications, reflecting their importance. Instances can be found in: microarray analysis [2, 3, 4, 5, 6], clinical prediction [7, 5, 6], text categorization [8, 9, 10, 11], image classification and face recognition [1], multi-label learning [12, 13], and classification of internet traffic [14]. A discussion on current challenges in feature selection, including online feature selection and the connection between feature selection and deep learning, is presented in [15].

Feature selection techniques can be categorized as classifier-dependent (*wrapper* and *embedded* methods) and classifier-independent (*filter* methods). Wrapper methods [16] search the space of feature subsets, using the classifier accuracy as the measure of utility for a candidate subset. There are clear disadvantages in using such approach. The computational cost is huge, while the selected features are specific for the considered classifier. Embedded methods [17, Ch. 5] exploit the structure of specific classes of classifiers to guide the feature selection process. In contrast, filter methods [17, Ch. 3] separate the classification and

ACCEPTED MANUSCRIPT

feature selection procedures, and define a heuristic ranking criterion that acts as a measure of the classification accuracy.

Filter methods differ among them in the way they quantify the benefits of including a particular feature in the set used in the classification process. Numerous heuristics have been suggested. Among these, methods for feature selection that rely on the concept of *mutual information* are the most popular. Mutual information (MI) captures linear and non-linear association between features, and is strongly related with the concept of *entropy*. Since considering the complete set of candidate features is too complex, filter methods usually operate sequentially and in the forward direction, adding one candidate feature at a time to the set of selected features. Here, the selected feature is the one that, among the set of candidate features, maximizes an objective function expressing the contribution of the candidate to the explanation of the class. A unifying approach for characterizing the different forward feature selection methods based on MI has been proposed by [18]. Moreover, [19] also provides an overview of the different feature selection methods, adding a list of open problems in the field.

Among the forward feature selection methods based on MI, the first proposed group [20, 21, 22, 23] is constituted by methods based on assumptions that were originally introduced in [20]. These methods attempt to select the candidate feature that leads to: maximum *relevance* between the candidate feature and the class; and minimum *redundancy* of the candidate feature with respect to the already selected features. Such redundancy, which we call *inter-feature redundancy*, is measured by the level of association between the candidate feature and the previously selected features. Considering inter-feature redundancy in the objective function is important, for instance, to avoid later problems of collinearity. In fact, selecting features that do not add value to the set of already selected ones in terms of class explanation, should be avoided.

A more recently proposed group of methods based on MI considers an additional term, resulting from the accommodation of possible dependencies between the features given the class [18]. This additional term is disregarded by

ACCEPTED MANUSCRIPT

the previous group of filter methods. Examples of methods from this second group are the ones proposed in: [24, 25, 26]. The additional term expresses the contribution of a candidate feature to the explanation of the class, when taken together with already selected features, which corresponds to a *class-relevant redundancy*. The effects captured by this type of redundancy are also called *complementarity* effects.

In this work we provide a comparison of forward feature selection methods based on mutual information using a theoretical framework. The framework is independent of specific datasets and classifiers and, therefore, provides a precise evaluation of the relative merits of the feature selection methods; it also allows unveiling several of their deficiencies. Our framework is grounded on the definition of a target (ideal) objective function and of a categorization of features according to their contribution to explanation of the class. We derive lower and upper bounds for the target objective function and establish a relation between these bounds and the feature types. The categorization of features has two novelties regarding previous works: we introduce the category of fully relevant features, features that fully explain the class together with already selected features, and we separate non-relevant features into irrelevant and redundant since, as we show, these categories have different properties regarding the feature selection process.

This framework provides a reference for evaluating and comparing actual feature selection methods. Actual methods are based on approximations of the target objective function, since the latter is difficult to estimate. We select a set of methods representative of the various types of approximations, and discuss the various drawbacks they introduced. Moreover, we analyze how each method copes with the good properties of the target objective function. Additionally, we define a distributional setting, based on a specific definition of class, features, and a novel performance metric; it provides a feature ranking for each method that is compared with the ideal feature ranking coming out of the theoretically framework. The setting was designed to challenge the actual feature selection methods, and illustrate the consequences of their drawbacks. Based on our

ACCEPTED MANUSCRIPT

work, we identify clearly the methods that should be avoided, and the methods that currently have the best performance.

Recently, there has been several attempts to undergo a theoretical evaluation of forward feature selection methods based on MI. For example, [18] and [19] provide an interpretation of the objective function of actual methods as approximations of a target objective function, which is similar to ours. However, they do not study the consequences of these approximations from a theoretical point-of-view, i.e. how the various types of approximations affect the good properties of the target objective function, which is the main contribution of our work. Moreover, they do not cover all types of feature selection methods currently proposed. [27] evaluated methods based on a distributional setting similar to ours, but the analysis is restricted to the group of methods that ignore complementarity, and again, does not address the theoretical properties of the methods. Table 1 summarizes the main contributions of the most relevant surveys on this topic and contrasts those with the work developed in this paper.

The rest of the paper is organized as follows. We introduce some background on entropy and MI in Section 2. This is followed, in Section 3, by the presentation of the main concepts associated with conditional MI and MI between three random vectors. In Section 4, we focus on explaining the general context concerning forward feature selection methods based on MI, namely the target objective function, the categorization of features, and the relation between the feature types and the bounds of the target objective function. In Section 5, we introduce representative feature selection methods based on MI, along with their properties and drawbacks. In Section 6, we present a distribution based setting where some of the main drawbacks of the representative methods are illustrated, using the minimum Bayes risk as performance evaluation measure to assess the quality of the methods. The main conclusions can be found in Section 7.

To facilitate the reading, we introduced in Table 2 the notation used in the paper.

ACCEPTED MANUSCRIPT

Table 1: Main contributions of selected surveys on feature selection methods.

Reference	Main Contributions
Guyon & Elisseeff (2003)[28]	Discuss objective functions, feature ranking, search methods, and validity assessment
Guyon et al. (2008)[17]	Part I of the book provides theoretical foundations to understand several families of feature selection methods
Brown et al. (2012)[18]	Unifying approach, relying on conditional likelihood, for characterizing different forward feature selection methods based on MI
Vergara & Estévez (2014)[19]	Provide an overview of the different feature selection methods, adding a list of open problems in the field
Pascoal et al. (2017)[27]	Evaluate methods, that ignore complementarity, based on a distributional setting
Present Work	Provide an interpretation of the objective function of known methods as approximations of a target objective function. Proved, from a theoretical point-of-view, how the various types of approximations affect the good properties of the target objective function, highlighting deficiencies of the chosen feature selection methods. The methods that should be avoided and the ones that currently have the best performance are identified

2. Entropy and mutual information

In this section, we present the main ideas behind the concepts of entropy and mutual information, along with their basic properties. In what follows, $\mathcal{X}$ denotes the support of a random vector $\boldsymbol{X}$. Moreover, we assume the convention

ACCEPTED MANUSCRIPT

Table 2: Notation used along the paper.

Symbol	Description
$\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$	Support of a random vector $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$
$P(\cdot)$	Probability function of a given event
$H(\cdot)$	Entropy of a discrete random vector
$H(\cdot\|\cdot)$	Conditional entropy of a discrete random vector, rv
$h(\cdot)$	Differential entropy of an absolutely continuous rv
$h(\cdot\|\cdot)$	Conditional differential entropy of absolutely continuous rv
$\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	n-dimensional multivariate normal distribution, with mean $\boldsymbol{\mu}$ $(n \times 1)$ and covariance matrix $\boldsymbol{\Sigma}$ $(n \times n)$
$\mathrm{MI}(\cdot, \cdot)$	Mutual Information, MI, between two rv
$\mathrm{MI}(\cdot, \cdot\|\cdot)$	Conditional MI between two rv given a third rv
$\mathrm{TMI}(\cdot, \cdot, \cdot)$	Triple MI between three rv
C	Class, which identifies the group each object belongs to
$\boldsymbol{S}$ ($\boldsymbol{F}$)	set of selected (unselected) features at a certain step of an iterative algorithm
$\mathrm{OF}(\cdot)$, $\mathrm{OF}'(\cdot)$	Objective functions to be optimize by the forward feature selection algorithms
$\|\boldsymbol{A}\|$	Number of elements of set $\boldsymbol{A}$
$\Gamma(\cdot)$	Approximating function

$0 \ln 0 = 0$, justified by continuity since $x \ln x \to 0$ as $x \to 0^+$.

2.1. Entropy

The concept of entropy [29] was initially motivated by problems in the field of telecommunications. Introduced for discrete random variables, the entropy is a measure of uncertainty. In the following, $P(A)$ denotes the probability of A.

Definition 1. *The entropy of a discrete random vector $\boldsymbol{X}$ is:*

$$H(\boldsymbol{X}) = -\sum_{\boldsymbol{x} \in \mathcal{X}} P(\boldsymbol{X} = \boldsymbol{x}) \ln P(\boldsymbol{X} = \boldsymbol{x}). \tag{1}$$

ACCEPTED MANUSCRIPT

Given an additional discrete random vector $\boldsymbol{Y}$*, the conditional entropy of* $\boldsymbol{X}$ *given* $\boldsymbol{Y}$ *is*

$$H(\boldsymbol{X}|\boldsymbol{Y}) = -\sum_{\boldsymbol{y}\in\mathcal{Y}}\sum_{\boldsymbol{x}\in\mathcal{X}} P(\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{Y}=\boldsymbol{y})P(\boldsymbol{Y}=\boldsymbol{y})\ln P(\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{Y}=\boldsymbol{y}).$$

Note that the entropy of $\boldsymbol{X}$ does not depend on the particular values taken by the random vector but only on the corresponding probabilities. It is clear that entropy is non-negative since each term of the summation in (1) is non-positive. Additionally, the value 0 is only obtained for a degenerate random variable.

An important property that results from Definition 1 is the so-called *chain rule* [30, Ch. 2]:

$$H(\boldsymbol{X}_1,...,\boldsymbol{X}_n) = \sum_{i=2}^{n} H(\boldsymbol{X}_i|\boldsymbol{X}_{i-1},...,\boldsymbol{X}_1) + H(\boldsymbol{X}_1), \tag{2}$$

where a sequence of random vectors, such as $(\boldsymbol{X}_1,...,\boldsymbol{X}_n)$ and $(\boldsymbol{X}_{i-1},...,\boldsymbol{X}_1)$ above, should be seen as the random vector that results from the concatenation of its elements.

2.2. Differential entropy

A logical way to adapt the definition of entropy to the case where we deal with an absolutely continuous random vector is to replace the probability (mass) function of a discrete random vector by the probability density function of an absolutely continuous random vector, as next presented. The resulting concept is called *differential entropy*. We let $f_{\boldsymbol{X}}$ denote the probability density function of an absolutely continuous random vector $\boldsymbol{X}$.

Definition 2. *The differential entropy of an absolutely continuous random vector* $\boldsymbol{X}$ *is:*

$$h(\boldsymbol{X}) = -\int_{\boldsymbol{x}\in\mathcal{X}} f_{\boldsymbol{X}}(\boldsymbol{x})\ln f_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x}. \tag{3}$$

Given an additional absolutely continuous random vector $\boldsymbol{Y}$*, such that* $(\boldsymbol{X},\boldsymbol{Y})$ *is also absolutely continuous, the conditional differential entropy of* $\boldsymbol{X}$ *given* $\boldsymbol{Y}$ *is*

$$h(\boldsymbol{X}|\boldsymbol{Y}) = -\int_{\boldsymbol{y}\in\mathcal{Y}} f_{\boldsymbol{Y}}(\boldsymbol{y})\int_{\boldsymbol{x}\in\mathcal{X}} f_{\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}}(\boldsymbol{x})\ln f_{\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}}(\boldsymbol{x})d\boldsymbol{x}\,d\boldsymbol{y}.$$

ACCEPTED MANUSCRIPT

It can be proved [30, Ch. 9] that the chain rule (2) still holds replacing entropy by differential entropy.

The notation that is used for differential entropy, h, is different from the notation used for entropy, H. This is justified by the fact that entropy and differential entropy do not share the same properties. For instance, non-negativity does not necessarily hold for differential entropy. Also note that $h(\boldsymbol{X}, \boldsymbol{X})$ and $h(\boldsymbol{X}|\boldsymbol{X})$ are not defined given that the pair $(\boldsymbol{X}, \boldsymbol{X})$ is not absolutely continuous. Therefore, relations involving entropy and differential entropy need to be interpreted in a different way.

Example 1. *If $\boldsymbol{X}$ is a random vector, of dimension n, following a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the value of the corresponding differential entropy is $\frac{1}{2}\ln\left((2\pi e)^n|\boldsymbol{\Sigma}|\right)$ [30, Ch. 9], where $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$. In particular, for the one-dimensional case, $X \sim \mathcal{N}(\mu, \sigma^2)$, the differential entropy is negative if $\sigma^2 < 1/2\pi e$, positive if $\sigma^2 > 1/2\pi e$, and zero if $\sigma^2 = 1/2\pi e$. Thus, a zero differential entropy does not have the same interpretation as in the discrete case. Moreover, the differential entropy can take arbitrary negative values.*

In the rest of the paper, when the context is clear, we will refer to differential entropy simply as entropy.

2.3. Mutual information

We now introduce mutual information (MI), which is a very important measure since it measures both linear and non-linear associations between random vectors.

2.3.1. Discrete case

Definition 3. *The MI between two discrete random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ is:*

$$\mathrm{MI}(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{\boldsymbol{x}\in\mathcal{X}} \sum_{\boldsymbol{y}\in\mathcal{Y}} P(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}) \ln \frac{P(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y})}{P(\boldsymbol{X} = \boldsymbol{x})P(\boldsymbol{Y} = \boldsymbol{y})}.$$

ACCEPTED MANUSCRIPT

MI satisfies the following [cf. 30, Ch. 9]:

$$\mathrm{MI}(\boldsymbol{X},\boldsymbol{Y}) = H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{Y}); \tag{4}$$

$$\mathrm{MI}(\boldsymbol{X},\boldsymbol{Y}) \geq 0; \tag{5}$$

$$\mathrm{MI}(\boldsymbol{X},\boldsymbol{X}) = H(\boldsymbol{X}). \tag{6}$$

Equality holds in (5) if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent random vectors.

According to (4), $\mathrm{MI}(\boldsymbol{X},\boldsymbol{Y})$ can be interpreted as the reduction in the uncertainty of $\boldsymbol{X}$ due to the knowledge of $\boldsymbol{Y}$. Note that, applying (2), we also have

$$\mathrm{MI}(\boldsymbol{X},\boldsymbol{Y}) = H(\boldsymbol{X}) + H(\boldsymbol{Y}) - H(\boldsymbol{X},\boldsymbol{Y}). \tag{7}$$

Another important property that immediately follows from (4) is

$$\mathrm{MI}(\boldsymbol{X},\boldsymbol{Y}) \leq \min(H(\boldsymbol{X}), H(\boldsymbol{Y})). \tag{8}$$

In sequence, in view of (4) and (5), we can conclude that, for any random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$H(\boldsymbol{X}|\boldsymbol{Y}) \leq H(\boldsymbol{X}). \tag{9}$$

This result is again coherent with the intuition that entropy measures uncertainty. In fact, if more information is added, about $\boldsymbol{Y}$ in this case, the uncertainty about $\boldsymbol{X}$ will not increase.

2.3.2. Continuous case

Definition 4. *The MI between two absolutely continuous random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, such that $(\boldsymbol{X},\boldsymbol{Y})$ is also absolutely continuous, is:*

$$\mathrm{MI}(\boldsymbol{X},\boldsymbol{Y}) = -\int_{\boldsymbol{y}\in\mathcal{Y}}\int_{\boldsymbol{x}\in\mathcal{X}} f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) \ln \frac{f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}{f_{\boldsymbol{X}}(\boldsymbol{x}) f_{\boldsymbol{Y}}(\boldsymbol{y})} d\boldsymbol{x}\, d\boldsymbol{y}.$$

It is straight-forward to check, given the similarities between this definition and Definition 3, that most properties from the discrete case still hold replacing entropy by differential entropy. In particular, the only property from (4) to (6) that cannot be restated for differential entropy is (6) since Definition 4 does not

ACCEPTED MANUSCRIPT

cover $\mathrm{MI}(\boldsymbol{X}, \boldsymbol{X})$, again because the pair $(\boldsymbol{X}, \boldsymbol{X})$ is not absolutely continuous. Additionally, restatements of (7) and (9) for differential entropy also hold.

On the whole, MI for absolutely continuous random vectors verifies the most important properties that are valid for the discrete case, including being symmetric in the arguments and non-negative. Moreover, the value 0 is obtained if and only if the involved random vectors are independent. However, a parallel of (8) for absolutely continuous random vectors, i.e. $\mathrm{MI}(\boldsymbol{X}, \boldsymbol{Y}) \leq \min(h(\boldsymbol{X}), h(\boldsymbol{Y}))$ may not hold. In fact, since $\mathrm{MI}(\boldsymbol{X}, \boldsymbol{Y}) \geq 0$ it follows that $\mathrm{MI}(\boldsymbol{X}, \boldsymbol{Y}) > \min(h(\boldsymbol{X}), h(\boldsymbol{Y}))$ whenever $h(\boldsymbol{X}) < 0$, a condition that holds for particular continuous random vectors $\boldsymbol{X}$, as shown in Example 1.

2.3.3. Combination of continuous with discrete random vectors

The definition of MI when we have an absolutely continuous random vector and a discrete random vector is also important in later stages of this article. For this reason, and despite the fact that the results that follow are naturally obtained from those that involve only either discrete or absolutely continuous vectors, we briefly go through them now.

Definition 5. *The MI between an absolutely continuous random vector $\boldsymbol{X}$ and a discrete random vector $\boldsymbol{Y}$ is given by either of the following two expressions:*

$$\begin{aligned}\mathrm{MI}(\boldsymbol{X}, \boldsymbol{Y}) &= \sum_{\boldsymbol{y} \in \mathcal{Y}} P(\boldsymbol{Y} = \boldsymbol{y}) \int_{\boldsymbol{x} \in \mathcal{X}} f_{\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}}(\boldsymbol{x}) \ln \frac{f_{\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}}(\boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})} d\boldsymbol{x} \\ &= \int_{\boldsymbol{x} \in \mathcal{X}} f_{\boldsymbol{X}}(\boldsymbol{x}) \sum_{\boldsymbol{y} \in \mathcal{Y}} P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}) \ln \frac{P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x})}{P(\boldsymbol{Y} = \boldsymbol{y})} d\boldsymbol{x}.\end{aligned}$$

The majority of the properties stated for the discrete case are still valid in this case. In particular, analogues of (4) hold, both in terms of entropies as well as in terms of differential entropies:

$$\mathrm{MI}(\boldsymbol{X}, \boldsymbol{Y}) = h(\boldsymbol{X}) - h(\boldsymbol{X}|\boldsymbol{Y}) \tag{10}$$

$$= H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X}). \tag{11}$$

Furthermore, $\mathrm{MI}(\boldsymbol{X}, \boldsymbol{Y}) \leq H(\boldsymbol{Y})$ is the analogue of (8) for this setting. Note

ACCEPTED MANUSCRIPT

that (11), but not (10), can be used to obtain an upper bound for $\mathrm{MI}(\boldsymbol{X}, \boldsymbol{Y})$ since $h(\boldsymbol{X}|\boldsymbol{Y})$ may be negative.

3. Triple mutual information and conditional mutual information

In this section, we discuss definitions and important properties associated with conditional MI and MI between three random vectors. Random vectors are considered to be discrete in this section as the generalization of the results for absolutely continuous random vectors would follow a similar approach.

3.1. Conditional mutual information

Conditional MI is defined in terms of entropies as follows, in a similar way to property (4) [cf. 26, 31].

Definition 6. *The conditional MI between two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ given the random vector $\boldsymbol{Z}$ is written as*

$$\mathrm{MI}(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}) = H(\boldsymbol{X}|\boldsymbol{Z}) - H(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{Z}). \tag{12}$$

Using (12) and an analogue of the chain rule for conditional entropy, we conclude that:

$$\mathrm{MI}(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}) = H(\boldsymbol{X}|\boldsymbol{Z}) + H(\boldsymbol{Y}|\boldsymbol{Z}) - H(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}). \tag{13}$$

In view of Definition 6, developing the involved terms according to Definition 3, we obtain:

$$\mathrm{MI}(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}) = E_{\boldsymbol{Z}}[\mathrm{MI}(\tilde{\boldsymbol{X}}(\boldsymbol{Z}), \tilde{\boldsymbol{Y}}(\boldsymbol{Z})], \tag{14}$$

where, for $\boldsymbol{z} \in \mathcal{Z}$, $(\tilde{\boldsymbol{X}}(\boldsymbol{z}), \tilde{\boldsymbol{Y}}(\boldsymbol{z}))$ is equal in distribution to $(\boldsymbol{X}, \boldsymbol{Y})|\boldsymbol{Z} = \boldsymbol{z}$.

Taking (5) and (14) into account,

$$\mathrm{MI}(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}) \geq 0, \tag{15}$$

and $\mathrm{MI}(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}) = 0$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are conditionally independent given $\boldsymbol{Z}$.

Moreover, from (12) and (15), we conclude the following result similar to (9):

$$H(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{Z}) \leq H(\boldsymbol{X}|\boldsymbol{Z}). \tag{16}$$

ACCEPTED MANUSCRIPT

3.2. Triple mutual information

The generalization of the concept of MI to more than two random vectors is not unique. One such generalization, associated with the concept of total correlation, was proposed by [32]. An alternative one, proposed by [33], is called *triple MI* (TMI). We will consider the latter since it is the most meaningful one in the context of objective functions associated with the problem of forward feature selection.

Definition 7. *The triple MI between three random vectors* $\boldsymbol{X}$*,* $\boldsymbol{Y}$*, and* $\boldsymbol{Z}$ *is defined as*

$$\mathrm{TMI}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = \sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{\boldsymbol{y}\in\mathcal{Y}}\sum_{\boldsymbol{z}\in\mathcal{Z}} P(\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{Z}=\boldsymbol{z})\times \ln\frac{P(\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y})P(\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{Z}=\boldsymbol{z})P(\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Z}=\boldsymbol{z})}{P(\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{Z}=\boldsymbol{z})P(\boldsymbol{X}=\boldsymbol{x})P(\boldsymbol{Y}=\boldsymbol{y})P(\boldsymbol{Z}=\boldsymbol{z})}.$$

Using the definition of MI and TMI, we can conclude that TMI and conditional MI are related in the following way, which provides extra intuition about the two concepts:

$$\mathrm{TMI}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = \mathrm{MI}(\boldsymbol{X},\boldsymbol{Y}) - \mathrm{MI}(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z}). \tag{17}$$

The TMI is not necessarily non-negative. This fact is exemplified and discussed in detail in the next section.

4. The forward feature selection problem

In this section, we focus on explaining the general context concerning forward feature selection methods based on mutual information. We first introduce target objective functions to be maximized in each step; we then define important concepts and prove some properties of such target objective functions. In the rest of this section, features are considered to be discrete for simplicity. The name target objective functions comes from the fact that, as we will argue, these are objective functions that perform exactly as we would desire ideally, so that a good method should reproduce its properties as well as possible.

ACCEPTED MANUSCRIPT

4.1. Target objective functions

Let C denote the class-variable, which identifies the group each object belongs to and $\boldsymbol{S}$ ($\boldsymbol{F}$) the set of selected (unselected) features at a given step of the iterative algorithm. Being so, $\boldsymbol{S} \cap \boldsymbol{F} = \emptyset$ and $\boldsymbol{S} \cup \boldsymbol{F}$ is the set with all input features. In what follows, when a set of random variables is in the argument of an entropy or MI term, it stands for the random vector composed by the random variables it contains. For example, $\mathrm{MI}(C, \{X_1, X_2\} \cup \{X_3\}) = \mathrm{MI}(C, \boldsymbol{X})$, where $\boldsymbol{X} = (X_1, X_2, X_3)$.

Given the set of selected features, forward feature selection methods aim to select a candidate feature $X_j \in \boldsymbol{F}$ such that

$$X_j = \arg\max_{X_i \in \boldsymbol{F}} \mathrm{MI}(C, \boldsymbol{S} \cup \{X_i\}).$$

Therefore, X_j is, among the features in $\boldsymbol{F}$, the feature X_i for which $\boldsymbol{S} \cup \{X_i\}$ maximizes the association (measured using MI) with the class, C. Note that we choose the feature that maximizes $\mathrm{MI}(C, X_i)$ in the first step (i.e., when $\boldsymbol{S} = \emptyset$).

Since $\mathrm{MI}(C, \boldsymbol{S} \cup \{X_i\}) = \mathrm{MI}(C, \boldsymbol{S}) + \mathrm{MI}(C, X_i|\boldsymbol{S})$ [cf. 22], in view of (17), the objective function evaluated at the candidate feature X_i can be written as

$$\begin{aligned} \mathrm{OF}(X_i) &= \mathrm{MI}(C, \boldsymbol{S}) + \mathrm{MI}(C, X_i|\boldsymbol{S}) \\ &= \mathrm{MI}(C, \boldsymbol{S}) + \mathrm{MI}(C, X_i) - \mathrm{TMI}(C, X_i, \boldsymbol{S}) \\ &= \mathrm{MI}(C, \boldsymbol{S}) + \mathrm{MI}(C, X_i) - \mathrm{MI}(X_i, \boldsymbol{S}) + \mathrm{MI}(X_i, \boldsymbol{S}|C). \end{aligned} \tag{18}$$

The feature selection methods try to approximate this objective function. However, since the term $\mathrm{MI}(C, \boldsymbol{S})$ does not depend on X_i, most approximations can be studied taking as a reference the simplified form of objective function given by

$$\mathrm{OF}'(X_i) = \mathrm{MI}(C, X_i) - \mathrm{MI}(X_i, \boldsymbol{S}) + \mathrm{MI}(X_i, \boldsymbol{S}|C).$$

This objective function has distinct properties from those of (18) and, therefore, deserves being addressed separately. Moreover, it is the reference objective function for most feature selection methods.

ACCEPTED MANUSCRIPT

The objective functions OF and OF′ can be written in terms of entropies, which provides a useful interpretation. Using (4) and (12), we obtain for the first objective function:

$$\mathrm{OF}(X_i) = H(C) - H(C|X_i, \boldsymbol{S}). \tag{19}$$

Maximizing $H(C) - H(C|X_i, \boldsymbol{S})$ provides the same candidate feature X_j as minimizing $H(C|X_i, \boldsymbol{S})$, for $X_i \in \boldsymbol{F}$. This means that the feature to be selected is the one leading to the minimal uncertainty of the class among the candidate features. As for the second objective function, we obtain, using (4), (19) and the fact that $\mathrm{OF}'(X_i) = \mathrm{OF}(X_i) - \mathrm{MI}(C, \boldsymbol{S})$:

$$\mathrm{OF}'(X_i) = H(C|\boldsymbol{S}) - H(C|X_i, \boldsymbol{S}). \tag{20}$$

This emphasizes that a feature that maximizes (19) also maximizes (20). In fact, the term that depends on X_i is the same in the two expressions.

We now provide bounds for the target objective functions.

Theorem 1. *Given a general candidate feature X_i:*

1. $H(C) - H(C|\boldsymbol{S}) \leq \mathrm{OF}(X_i) \leq H(C)$.
2. $0 \leq \mathrm{OF}'(X_i) \leq H(C|\boldsymbol{S})$.

Proof. Using the representations (19) and (20) for $\mathrm{OF}(X_i)$ and $\mathrm{OF}'(X_i)$, respectively, the upper bounds in statements 1 and 2 follow from the fact that $H(C|X_i, \boldsymbol{S}) \geq 0$. As for the lower bounds, in the case of statement 2, it comes directly from the fact that $\mathrm{OF}'(X_i) = \mathrm{MI}(C, X_i|\boldsymbol{S}) \geq 0$, in view of (20) and (12). As for statement 1, given that, from (19) and (12), $\mathrm{OF}(X_i) = H(C) - H(C|X_i, \boldsymbol{S}) = H(C) - H(C|\boldsymbol{S}) + \mathrm{MI}(C, X_i|\boldsymbol{S})$, we again only need to use the fact that $\mathrm{MI}(C, X_i|\boldsymbol{S}) \geq 0$, to conclude that $\mathrm{OF}(X_i) \geq H(C) - H(C|\boldsymbol{S})$. □

The upper bound for OF, $H(C)$, corresponds to the uncertainty in C, and the upper bound on OF′, $H(C|\boldsymbol{S})$, corresponds to the uncertainty in C not explained by the already selected features, $\boldsymbol{S}$. This is coherent with the fact that OF′ ignores the term $\mathrm{MI}(C, \boldsymbol{S})$. The lower bound for OF corresponds to the uncertainty in C already explained by $\boldsymbol{S}$.

ACCEPTED MANUSCRIPT

4.2. Feature types and their properties

Features can be characterized according to their usefulness in explaining the class at a particular step of the feature selection process. We next define four types of features: irrelevant, redundant, relevant, and fully relevant.

Definition 8. *Given a subset of already selected features, $\boldsymbol{S}$, at a certain step of a forward sequential method, where the class is C, and a candidate feature X_i, then:*

- *X_i is irrelevant given $(C, \boldsymbol{S})$ if* $\mathrm{MI}(C, X_i|\boldsymbol{S}) = 0 \wedge H(X_i|\boldsymbol{S}) > 0$*;*
- *X_i is redundant given $\boldsymbol{S}$ if* $H(X_i|\boldsymbol{S}) = 0$*;*
- *X_i is relevant given $(C, \boldsymbol{S})$ if* $\mathrm{MI}(C, X_i|\boldsymbol{S}) > 0$*;*
- *X_i is fully relevant given $(C, \boldsymbol{S})$ if* $H(C|X_i, \boldsymbol{S}) = 0 \wedge H(C|\boldsymbol{S}) > 0$.

Relevant features given $(C, \boldsymbol{S})$ are features X_i that add information to the explanation of the class not provided by the set $\boldsymbol{S}$ already selected features, i.e. for which $\mathrm{MI}(C, X_i|\boldsymbol{S}) > 0$.

If $\boldsymbol{S} = \emptyset$, then $\mathrm{MI}(C, X_i|\boldsymbol{S})$*,* $H(X_i|\boldsymbol{S})$*,* $H(C|\boldsymbol{S})$*, and* $H(C|X_i, \boldsymbol{S})$ *should be replaced by* $\mathrm{MI}(C, X_i)$*,* $H(X_i)$*,* $H(C)$*, and* $H(C|X_i)$*, respectively.*

A feature X_i is fully relevant given $(C, \boldsymbol{S})$ if jointly with the already selected features completely explains the class, i.e. $H(C|X_i, \boldsymbol{S}) = 0$, but $\boldsymbol{S}$ alone is not enough to completely explain the class, i.e. $H(C|\boldsymbol{S}) > 0$. Note that fully relevant features given $(C, \boldsymbol{S})$ are relevant features given $(C, \boldsymbol{S})$ since the conditions $H(C|X_i, \boldsymbol{S}) = 0$ and $H(C|\boldsymbol{S}) > 0$ imply that

$$\mathrm{MI}(C, X_i|\boldsymbol{S}) = H(C|\boldsymbol{S}) - H(C|X_i, \boldsymbol{S}) > 0.$$

The features that are not relevant are either irrelevant or redundant. Accordingly, irrelevant and redundant features given $(C, \boldsymbol{S})$ are those that do not add information to the explanation of the class not provided by the already selected features, i.e. such that $\mathrm{MI}(C, X_i|\boldsymbol{S}) = 0$.

ACCEPTED MANUSCRIPT

A feature X_i is irrelevant given $(C, \boldsymbol{S})$ if not being completely explained by the already selected features, i.e. $H(X_i|\boldsymbol{S}) > 0$, it is independent of the class given the already selected features, i.e. $\mathrm{MI}(C, X_i|\boldsymbol{S}) = 0$.

Finally, X_i is redundant given $\boldsymbol{S}$ if X_i is completely explained by the already selected features, i.e. $H(X_i|\boldsymbol{S}) = 0$.

Under this definition, irrelevant, redundant, and relevant features form a partition of the set of candidate features $\boldsymbol{F}$.

Definition 8 introduces two novelties regarding previous works: first, we separate non-relevant features in two categories, of irrelevant and redundant features; second, we introduce the important category of fully relevant features. Note that, under this definition, irrelevant, redundant, and relevant features form a partition of the set of candidate features $\boldsymbol{F}$.

Our motivation for separating irrelevant from redundant features is that, while a redundant feature remains redundant at all subsequent steps of the feature selection process, the same does not hold necessarily for irrelevant features. The following example illustrates how an irrelevant feature can later become relevant.

Example 2. *We consider a class $C = (X + Y)^2$ where X and Y are two independent candidate features that follow uniform distributions on $\{-1, 1\}$. C follows a uniform distribution on $\{0, 4\}$ and, as a result, the entropies of X, Y and C are $\ln(2)$. It can be easily checked that both X and Y are independent of the class. In the feature selection process, both features are initially irrelevant since, due to their independence from C, $\mathrm{MI}(C, X) = \mathrm{MI}(C, Y) = 0$. Suppose that X is selected first. Then, Y becomes relevant since $\mathrm{MI}(C, Y|X) = \ln(2) > 0$, and it is even fully relevant since $H(C|Y, X) = 0$ and $H(C|X) = \ln(2) > 0$.*

The following theorem shows that redundant features always remain redundant.

Theorem 2. *If a feature is redundant given $\boldsymbol{S}$, then it is also redundant given $\boldsymbol{S}'$, for $\boldsymbol{S} \subset \boldsymbol{S}'$.*

ACCEPTED MANUSCRIPT

Proof. Suppose that X_i is a redundant feature given $\boldsymbol{S}$, so that $H(X_i|\boldsymbol{S}) = 0$, and $\boldsymbol{S} \subset \boldsymbol{S}'$. This implies that $H(X_i|\boldsymbol{S}') = 0$ by (16). As a result, X_i is also redundant given $\boldsymbol{S}'$. □

This result has an important practical consequence: features that are found redundant at a certain step of the feature selection process can be immediately removed from the set of candidate features $\boldsymbol{F}$, alleviating in this way the computational effort associated with the feature selection process.

Fully relevant features form an important subgroup of relevant features since, together with already selected features, they completely explain the class, i.e. $H(C|\boldsymbol{S})$ becomes 0 after selecting a fully relevant feature. Thus, after selecting a fully relevant feature, all remaining unselected features are necessarily either irrelevant or redundant and the algorithm must stop. This also means that detecting a fully relevant feature can be used as a stopping criterion of forward feature selection methods. The condition $H(C|\boldsymbol{S}) > 0$ in the definition of fully relevant feature is required since an unselected feature can no longer be considered of this type after $H(C|\boldsymbol{S}) = 0$.

A stronger condition that could be considered as a stopping criterion is $H(C|\boldsymbol{S}) = H(C|\boldsymbol{S}, \boldsymbol{F})$, meaning that the (complete) set of candidate features $\boldsymbol{F}$ has no further information to explain the class. As in the previous case, the candidate features will all be irrelevant or redundant. However, since forward feature selection algorithms only consider one candidate feature at each iteration, and the previous condition requires considering all candidate features simultaneously, such condition cannot be used as a stopping criterion.

Regarding the categorization of features introduced by other authors, [18] considered only one category of non-relevant features, named irrelevant, consisting of the candidate features X_i such that $\mathrm{MI}(C, X_i|\boldsymbol{S}) = 0$. [34] and [19] considered both irrelevant and redundant features. Their definition of irrelevant feature is the one of [18]; redundant features are defined as features such that $H(X_i|\boldsymbol{S}) = 0$. Since the latter condition implies that $\mathrm{MI}(C, X_i|\boldsymbol{S}) = 0$ by (4) and (16), it turns out that redundant features become a special case of irrelevant

ACCEPTED MANUSCRIPT

ones, which is not in agreement with our definition.

According to the feature types introduced above, a good feature selection method must select, at a given step, a relevant feature, preferably a fully relevant one, keep irrelevant features for future consideration, and discard redundant features. The following theorem relates these desirable properties with the values taken by the target objective functions.

Theorem 3.

1. *If X_i is a fully relevant feature given $(C, \boldsymbol{S})$, then* $\mathrm{OF}(X_i) = H(C)$ *and* $\mathrm{OF}'(X_i) = H(C|\boldsymbol{S})$*, i.e., the maximum possible values taken by the target objective functions are reached; recall Theorem 1.*
2. *If X_i is an irrelevant feature given $(C, \boldsymbol{S})$, then* $\mathrm{OF}(X_i) = H(C) - H(C|\boldsymbol{S})$ *and* $\mathrm{OF}'(X_i) = 0$*, i.e., the minimum possible values of the target objective functions are reached; recall Theorem 1.*
3. *If X_i is a redundant feature given $\boldsymbol{S}$, then* $\mathrm{OF}(X_i) = H(C) - H(C|\boldsymbol{S})$ *and* $\mathrm{OF}'(X_i) = 0$*, i.e., the minimum possible values of the target objective functions are reached; recall Theorem 1.*
4. *If X_i is a relevant feature, but not fully relevant, given $(C, \boldsymbol{S})$, then* $H(C) - H(C|\boldsymbol{S}) < \mathrm{OF}(X_i) < H(C)$ *and* $0 < \mathrm{OF}'(X_i) < H(C|\boldsymbol{S})$.

Proof. The two equalities in statement 1 are an immediate consequence of equations (19) and (20), using the fact that $H(C|X_i, \boldsymbol{S}) = 0$ if X_i is fully relevant given $(C, \boldsymbol{S})$.

Suppose that X_i is an irrelevant feature given $(C, \boldsymbol{S})$, so that $\mathrm{MI}(C, X_i|\boldsymbol{S}) = 0$. Then, the relation $\mathrm{OF}'(X_i) = 0$ results directly from the fact that $\mathrm{OF}'(X_i) = \mathrm{MI}(C, X_i|\boldsymbol{S})$, in view of (20) and (12). Conversely, the relation $\mathrm{OF}(X_i) = H(C) - H(C|\boldsymbol{S})$ follows from the fact that $\mathrm{OF}(X_i) = H(C) - H(C|\boldsymbol{S}) + \mathrm{MI}(C, X_i|\boldsymbol{S})$. As a result, statement 2 is verified.

The equalities in statement 3 follow likewise since $\mathrm{MI}(C, X_i|\boldsymbol{S}) = 0$ if X_i is a redundant feature given $\boldsymbol{S}$.

ACCEPTED MANUSCRIPT

As for statement 4, we need to prove that the objective functions neither take the minimum nor the maximum value for a relevant feature that is not fully relevant. We start by checking that the minimum values are not reached. The proof is similar to that of statement 2. Since $\mathrm{OF}'(X_i) = \mathrm{MI}(C, X_i|\boldsymbol{S})$, in view of (20) and (12), and since the assumption is that $\mathrm{MI}(C, X_i|\boldsymbol{S}) > 0$, then $\mathrm{OF}'(X_i)$ is surely larger than 0. Concerning $\mathrm{OF}(X_i)$, since $\mathrm{OF}(X_i) = H(C) - H(C|\boldsymbol{S}) + \mathrm{MI}(C, X_i|\boldsymbol{S})$ and $\mathrm{MI}(C, X_i|\boldsymbol{S}) > 0$, $\mathrm{OF}(X_i)$ must be larger than $H(C) - H(C|\boldsymbol{S})$. Concerning the upper bounds, the proof is now similar to that of statement 1. If the feature X_i is not fully relevant given $(C, \boldsymbol{S})$, meaning that $H(C|\boldsymbol{S}, X_i) > 0$, the desired conclusions immediately follow from (19) and (20). □

Thus, fully relevant (irrelevant and redundant) features achieve the maximum (minimum) of the objective functions, and relevant features that are not fully relevant achieve a value between the maximum and the minimum values of the objective functions. These properties assure that the ordering of features at a given step of the feature selection process is always correct. Note that irrelevant and redundant features can be discriminated by evaluating $H(X_i|\boldsymbol{S})$.

4.3. Complementarity

The concept of complementarity is associated with the term $\mathrm{TMI}(C, X_i, \boldsymbol{S})$ of the target objective function, as next introduced.

Definition 9. *Given a subset of already selected features, $\boldsymbol{S}$, at a certain step of a forward sequential method, where the class is C, and a candidate feature X_i, we say that X_i and $\boldsymbol{S}$ are complementary with respect to C if $-\mathrm{TMI}(C, X_i, \boldsymbol{S}) > 0$, where, as defined in* (17)*,*

$$\mathrm{TMI}(C, X_i, \boldsymbol{S}) = \mathrm{MI}(X_i, \boldsymbol{S}) - \mathrm{MI}(X_i, \boldsymbol{S}|C).$$

Moreover, following [24], we call $\mathrm{MI}(X_i, \boldsymbol{S})$ the inter-feature redundancy term, and, following [24, 35], we call $\mathrm{MI}(X_i, \boldsymbol{S}|C)$ class-relevant redundancy term.

ACCEPTED MANUSCRIPT

Authors in [31] provided an interesting interpretation of complementarity, noting that

$$-\mathrm{TMI}(C, X_i, \boldsymbol{S}) = \mathrm{MI}(\{X_i\} \cup \boldsymbol{S}, C) - \mathrm{MI}(X_i, C) - \mathrm{MI}(\boldsymbol{S}, C).$$

Thus, if $-\mathrm{TMI}(C, X_i, \boldsymbol{S}) > 0$, then $\mathrm{MI}(\{X_i\} \cup \boldsymbol{S}, C) > \mathrm{MI}(X_i, C) + \mathrm{MI}(\boldsymbol{S}, C)$. Therefore, $-\mathrm{TMI}(C, X_i, \boldsymbol{S}) > 0$ measures the gain resulting from considering X_i and $\boldsymbol{S}$ together, instead of considering them separately, when measuring the association with the class C.

Interestingly, [36] refer to complementarity as the existence of *positive interaction*, or *synergy*, between X_i and $\boldsymbol{S}$ with respect to C.

The inter-feature redundancy term, $\mathrm{MI}(X_i, \boldsymbol{S})$, measures the association between the candidate feature and the already selected features. Following [24], we call this term *inter-feature redundancy*. It is sometimes coined as the *bad* redundancy since it expresses the information of the candidate feature already contained in the set of already selected features.

Given that $\mathrm{MI}(X_i, \boldsymbol{S}) \geq 0$, a negative TMI is necessarily associated with a positive value of $\mathrm{MI}(X_i, \boldsymbol{S}|C)$. Thus, the class-relevant redundancy term, $\mathrm{MI}(X_i, \boldsymbol{S}|C)$, called conditional redundancy in [18], expresses the contribution of a candidate feature to the explanation of the class, when taken together with already selected features. Class-relevant redundancy is sometimes coined as the *good* redundancy since it expresses an association that contributes to the explanation of the class. Furthermore, [17] highlights that "correlation does not imply redundancy" to stress that association between X_i and $\boldsymbol{S}$ is not necessarily bad.

Note that TMI takes negative values whenever the class-relevant redundancy exceeds the inter-feature redundancy, i.e. $\mathrm{MI}(X_i, \boldsymbol{S}|C) > \mathrm{MI}(X_i, \boldsymbol{S})$. A candidate feature X_i for which $\mathrm{TMI}(C, X_i, \boldsymbol{S})$ is negative is a relevant feature, i.e. $\mathrm{MI}(C, X_i|\boldsymbol{S}) \geq 0$. Thus, a candidate feature may be relevant even if it is strongly associated with the already selected features. Moreover, class-relevant redundancy may turn a feature that was initially irrelevant into a relevant feature, as illustrated in Example 2. In that example, the candidate feature Y was in-

ACCEPTED MANUSCRIPT

dependent of the already selected one, X, i.e. $\mathrm{MI}(X_i, \boldsymbol{S}) = \mathrm{MI}(Y, X) = 0$, but Y taken together with X had a positive contribution to the explanation of the class (indeed it fully explained the class), since the class-relevant redundancy is positive, i.e. $\mathrm{MI}(X_i, \boldsymbol{S}|C) = \mathrm{MI}(Y, X|C) = \ln(2) > 0$.

5. Representative feature selection methods

The target objective functions discussed in Section 4 cannot be used in practice since they require the joint distribution of $(C, X_i, \boldsymbol{S})$, which is not known and has to be estimated. This becomes more and more difficult as the cardinality of $\boldsymbol{S}$, denoted by $|\boldsymbol{S}|$ from here on, increases.

The common solution is to use approximations, leading to different feature selection methods. For the analysis in this paper, we selected a set of methods representative of the main types of approximations to the target objective functions. In what follows, we first describe the representative methods, and discuss the advantages and drawbacks resulting from their underlying approximations; we then discuss how these methods cope with the desirable properties given by Theorem 1 and Theorem 3; finally, we briefly refer to other methods proposed in the literature and how they relate to the representative ones. In this section, features are considered to be discrete for simplicity.

5.1. The methods, their advantages and drawbacks

The methods selected to represent the main types of approximations to the target objective functions are: MIM [37], MIFS [20], mRMR [21], maxMIFS [27], CIFE [24], JMI [25], CMIM [26], and JMIM [38]. These methods are listed in Table 3, together with their objective functions. Note that, for all methods, including mRMR and JMI, the objective function in the first step of the algorithm is simply $\mathrm{MI}(C, X_i)$. This implies, in particular, that the first feature to be selected is the same in all methods.

The representations of the objective functions of CMIM and JMIM in the references where they were proposed [26, 38] differ from the ones in Table 3.

ACCEPTED MANUSCRIPT

More concretely, their objective functions were originally formalized in terms of minimum operators:

$$\mathrm{OF}_{\mathrm{CMIM}}(X_i) = \min_{X_s \in \boldsymbol{S}} \mathrm{MI}(C, X_i | X_s); \tag{21}$$

$$\mathrm{OF}_{\mathrm{JMIM}}(X_i) = \min_{X_s \in \boldsymbol{S}} \left\{ \mathrm{MI}(C, X_s) + \mathrm{MI}(C, X_i | X_s) \right\}. \tag{22}$$

The representations in Table 3 result from the above ones using simple algebraic manipulation; recall (17). They allow a nicer and unified interpretation of the objective functions. For instance, they allow noticing much more clearly the similarities between maxMIFS and CMIM, as well as between CMIM and JMIM.

The methods differ in the way their objective functions approximate the target objective functions. All methods except JMIM have objective functions that can be seen as approximations of the target OF'; the objective function of JMIM can be seen as an approximation of the target OF.

The methods can be characterized at two levels. The first level relates to the type of terms (relevancy and redundancy) that are included in the objective function. The second level relates to function used to approximate the redundancy terms. These approximations introduce drawbacks in the feature selection process with different degrees of severity. The various advantages and drawbacks are summarized in Table 4.

5.1.1. Types of terms included in the objective function

We start by noting that all selected methods include in their objective functions the term $\mathrm{MI}(C, X_i)$, accounting for the association between the candidate feature and the class.

However, MIM only includes this term. This method discards the whole TMI term of the target objective function OF', i.e.

$$\mathrm{OF}'(X_i) \approx \mathrm{MI}(C, X_i). \tag{23}$$

Thus, MIM ranks features accounting only for relevancy effects, and completely ignores feature redundancy.

ACCEPTED MANUSCRIPT

Table 3: Objective functions of the representative feature selection methods, evaluated at candidate feature X_i.

Method	Objective function evaluated at X_i
MIM	$\mathrm{MI}(C, X_i)$
MIFS	$\mathrm{MI}(C, X_i) - \beta \sum_{X_s \in \boldsymbol{S}} \mathrm{MI}(X_i, X_s)$
mRMR	$\mathrm{MI}(C, X_i) - \frac{1}{\lvert\boldsymbol{S}\rvert} \sum_{X_s \in \boldsymbol{S}} \mathrm{MI}(X_i, X_s)$
maxMIFS	$\mathrm{MI}(C, X_i) - \max_{X_s \in \boldsymbol{S}} \mathrm{MI}(X_i, X_s)$
CIFE	$\mathrm{MI}(C, X_i) - \sum_{X_s \in \boldsymbol{S}} (\mathrm{MI}(X_i, X_s) - \mathrm{MI}(X_i, X_s \lvert C))$
JMI	$\mathrm{MI}(C, X_i) - \frac{1}{\lvert\boldsymbol{S}\rvert} \sum_{X_s \in \boldsymbol{S}} (\mathrm{MI}(X_i, X_s) - \mathrm{MI}(X_i, X_s \lvert C))$
CMIM	$\mathrm{MI}(C, X_i) - \max_{X_s \in \boldsymbol{S}} \{\mathrm{MI}(X_i, X_s) - \mathrm{MI}(X_i, X_s \lvert C)\}$
JMIM	$\mathrm{MI}(C, X_i) - \max_{X_s \in \boldsymbol{S}} \{\mathrm{MI}(X_i, X_s) - \mathrm{MI}(X_i, X_s \lvert C) - \mathrm{MI}(C, X_s)\}$

Table 4: Advantages and drawbacks of the representative feature selection methods.

Advantage	MIM	MIFS	mRMR	maxMIFS	CIFE	JMI	CMIM	JMIM
Class association	X	X	X	X	X	X	X	X
Feature redundancy		X	X	X	X	X	X	X
Feature complementarity					X	X	X	X
Drawback	**MIM**	**MIFS**	**mRMR**	**maxMIFS**	**CIFE**	**JMI**	**CMIM**	**JMIM**
Redundancy undervalued			X	X		X	X	X
Redundancy overscaled		X			X			
Complementarity penalized							X	X
Unimportant term approximated								X

ACCEPTED MANUSCRIPT

All other methods account for some type of redundancy effects. However, methods MIFS, mRMR, and maxMIFS ignore feature complementarity effects, since they only include the inter-feature redundancy term, while methods CIFE, JMI, CMIM, and JMIM account for complementarity effects, since they include both the inter-feature and class-relevant redundancy terms.

Methods ignoring feature complementarity effects Methods MIFS, mRMR, and maxMIFS discard the class-relevant redundancy term, i.e. they ignore complementarity effects, and consider an approximation to the inter-feature redundancy term of OF′ , i.e.

$$\mathrm{OF}'(X_i) \approx \mathrm{MI}(C, X_i) - \Gamma(\mathrm{MI}(X_i, \boldsymbol{S}). \tag{24}$$

where Γ denotes an approximating function. In this case, the TMI can no longer take negative values, since it reduces to the term $\mathrm{MI}(X_i, \boldsymbol{S})$. As discussed in Section 4.3, the complementarity expresses the contribution of a candidate feature to the explanation of the class, when taken together with already selected features, and ignoring this contribution may lead to gross errors in the feature selection process. This drawback was noted by [18].

Methods considering feature complementarity effects Methods CIFE, JMI, CMIM, and JMIM include both the inter-feature and class-relevant redundancy terms, but differ on the underlying objective function they try to approximate. Specifically, for CIFE, JMI, and CMIM,

$$\mathrm{OF}'(X_i) \approx \mathrm{MI}(C, X_i) - \Gamma(\mathrm{TMI}(C, X_i, \boldsymbol{S}))$$

and for JMIM,

$$\mathrm{OF}(X_i) \approx \mathrm{MI}(C, X_i) - \Gamma(\mathrm{TMI}(C, X_i, \boldsymbol{S}) - \mathrm{MI}(C, \boldsymbol{S})),$$

where again Γ denotes an approximating function.

5.1.2. Approximating functions of redundancy terms

The second level of characterization of feature selection methods relates to the way the *surviving* redundancy terms are approximated. In all cases,

ACCEPTED MANUSCRIPT

the terms of the objective functions that depend on the set $\boldsymbol{S}$, i.e. $\mathrm{MI}(C, \boldsymbol{S})$, $\mathrm{MI}(X_i, \boldsymbol{S})$, and $\mathrm{MI}(X_i, \boldsymbol{S}|C)$, which are difficult to estimate, are approximated through a function of the already selected features X_s, $X_s \in \boldsymbol{S}$, taken individually. Hereafter, we denote an already selected feature $X_s \in \boldsymbol{S}$ simply by X_s. Considering only individual associations neglects higher order associations, i.e. between a candidate and two or more already selected features. There are three types of approximating functions: (i) a sum of X_s terms scaled by a constant, (ii) an average of X_s terms, and (iii) a maximization over X_s terms.

Approximation through a sum of X_s terms scaled by a constant The first type of approximation is used by MIFS and CIFE. MIFS approximates the TMI (inter-feature redundancy only) through a sum of terms scaled by a constant β. CIFE approximates the TMI (both inter-feature and class-relevant redundancies) by a sum of X_s terms, i.e.

$$\begin{aligned}\mathrm{TMI}(C, X_i, \boldsymbol{S}) \approx \sum_{X_s \in \boldsymbol{S}} \mathrm{TMI}(C, X_i, X_s) &= \sum_{X_s \in \boldsymbol{S}} [\mathrm{MI}(X_i, X_s) - \mathrm{MI}(X_i, X_s|C)] \\ &= \sum_{X_s \in \boldsymbol{S}} \mathrm{MI}(X_i, X_s) - \sum_{X_s \in \boldsymbol{S}} \mathrm{MI}(X_i, X_s|C).\end{aligned}$$

In both cases, a problem arises because the TMI is approximated by a sum of terms which individually have the same scale as the term they try to approximate. This results in an approximation of the TMI that has a larger scale than the original term. Since these terms are both redundancy terms, we will refer to this as the *redundancy overscaled* drawback. It becomes more and more severe as $\boldsymbol{S}$ grows. This drawback was also noted by [18], referring to it as the problem of not balancing the magnitudes of the relevancy and the redundancy.

Approximation through an average of X_s terms The second type of approximation was introduced to overcome the redundancy overscaled drawback. It is used by mRMR and JMI, and replaces the TMI by an average of X_s terms.

ACCEPTED MANUSCRIPT

In particular, for JMI,

$$\begin{aligned}\mathrm{TMI}(C, X_i, \boldsymbol{S}) \approx & \frac{1}{|\boldsymbol{S}|} \sum_{X_s \in \boldsymbol{S}} \mathrm{TMI}(C, X_i, X_s) \\ = & \frac{1}{|\boldsymbol{S}|} \sum_{X_s \in \boldsymbol{S}} [\mathrm{MI}(X_i, X_s) - \mathrm{MI}(X_i, X_s|C)] \\ = & \frac{1}{|\boldsymbol{S}|} \sum_{X_s \in \boldsymbol{S}} \mathrm{MI}(X_i, X_s) - \frac{1}{|\boldsymbol{S}|} \sum_{X_s \in \boldsymbol{S}} \mathrm{MI}(X_i, X_s|C).\end{aligned}$$

The mRMR approximation is similar, but without the class-relevant redundancy terms. This approximation solves the overscaling problem but introduces another drawback. In fact, since $\mathrm{MI}(X_i, \boldsymbol{S}) \geq \mathrm{MI}(X_i, X_s)$, implying that $\mathrm{MI}(X_i, \boldsymbol{S}) \geq \frac{1}{|\boldsymbol{S}|} \sum_{X_s \in \boldsymbol{S}} \mathrm{MI}(X_i, X_s)$, the approximation undervalues the inter-feature redundancy; at the same time, given that $\mathrm{MI}(X_i, \boldsymbol{S}|C) \geq \mathrm{MI}(X_i, X_s|C)$, implying $\mathrm{MI}(X_i, \boldsymbol{S}|C) \geq \frac{1}{|\boldsymbol{S}|} \sum_{X_s \in \boldsymbol{S}} \mathrm{MI}(X_i, X_s|C)$, it also undervalues the class-relevant redundancy. We call this drawback *redundancy undervalued*.

Approximation through a maximization over X_s terms The third type of approximation was also introduced to overcome the redundancy overscaled drawback, and is a maximization over X_s terms. This approximation is used differently in maxMIFS and CMIM, on one side, and JMIM, on the other. Methods maxMIFS and CMIM just replace the TMI by a maximization over X_s terms. In particular, for CMIM,

$$\mathrm{TMI}(C, X_i, \boldsymbol{S}) \approx \max_{X_s \in \boldsymbol{S}} \mathrm{TMI}(C, X_i, X_s) = \max_{X_s \in \boldsymbol{S}} \left(\mathrm{MI}(X_i, X_s) - \mathrm{MI}(X_i, X_s|C)\right).$$

The maxMIFS approximation is similar, but without the class-relevant redundancy terms. The discussion regarding the quality of the approximation is more complex in this case.

maxMIFS We start by maxMIFS. In this case, since $\mathrm{MI}(X_i, \boldsymbol{S}) \geq \mathrm{MI}(X_i, X_s)$,

$$\mathrm{MI}(X_i, \boldsymbol{S}) \geq \max_{X_s \in \boldsymbol{S}} \mathrm{MI}(X_i, X_s) \geq \frac{1}{|\boldsymbol{S}|} \sum_{X_s \in \boldsymbol{S}} \mathrm{MI}(X_i, X_s). \tag{25}$$

Thus, maxMIFS approximation still undervalues the inter-feature redundancy, but is clearly better than the one considering an average. Indeed, maximizing

ACCEPTED MANUSCRIPT

over X_s is the best possible approximation, under the restriction that only one X_s is considered.

CMIM Regarding CMIM, we first note that a relationship similar to (25) also holds for the class-relevant redundancy, i.e.

$$\mathrm{MI}(X_i, \boldsymbol{S}|C) \geq \max_{X_s \in \boldsymbol{S}} \mathrm{MI}(X_i, X_s|C) \geq \frac{1}{|\boldsymbol{S}|} \sum_{X_s \in \boldsymbol{S}} \mathrm{MI}(X_i, X_s|C),$$

since $\mathrm{MI}(X_i, \boldsymbol{S}|C) \geq \mathrm{MI}(X_i, X_s|C)$. However, while it is true for the two individual terms that compose the TMI that $\mathrm{MI}(X_i, \boldsymbol{S}) \geq \max_{X_s \in \boldsymbol{S}} \mathrm{MI}(X_i, X_s)$ and $\mathrm{MI}(X_i, \boldsymbol{S}|C) \geq \max_{X_s \in \boldsymbol{S}} \mathrm{MI}(X_i, X_s|C)$, it is no longer true that $\mathrm{TMI}(C, X_i, \boldsymbol{S}) \geq \max_{X_s \in \boldsymbol{S}} \mathrm{TMI}(C, X_i, X_s)$. Thus, the maximization over X_s terms of CMIM is not as effective as that of maxMIFS. Moreover, applying a maximization jointly to the difference between the inter-feature and the class-relevant redundancy terms clearly favors X_s features that together with X_i have small class-relevant redundancy, i.e. a small value of $\mathrm{MI}(X_i, X_s|C)$. This goes against the initial purpose of methods that, like CMIM, introduced complementarity effects in forward feature selection methods. We call this drawback *complementarity penalized*. We now give an example that illustrates how this drawback may impact the feature selection process.

Example 3. *Assume that we have the same features as in Example 2, plus two extra features W and Z, independent of any vector containing other random variables of the set $\{W, Z, X, Y, C\}$. Moreover, consider the objective function of CMIM.*

In the first step, the objective function value is 0 *for all features. We assume that W is selected first. In this case, at the second step, the objective functions value is again* 0 *for all features. We assume that X is selected. At the third step, Y should be selected since it is fully relevant and Z is irrelevant. At this step, the objective function value at Z is 0. The objective function at Y requires a closer attention. Since Y is independent of the class, $\mathrm{MI}(Y, C) = 0$, the target*

ACCEPTED MANUSCRIPT

objective function evaluated at Y is

$$-\mathrm{TMI}(C,Y,\{W,X\}) = -\left[\mathrm{MI}(Y,\{W,X\}) - \mathrm{MI}(Y,\{W,X\}|C)\right]$$
$$= -\left[0 - (H(Y|C) - H(Y|W,X,C))\right] = -(0 - \ln(2)) = \ln(2),$$

and the objective function of CMIM evaluated at Y is

$$-\max\{\mathrm{TMI}(C,Y,W), \mathrm{TMI}(C,Y,X)\}$$
$$= -\max\{\mathrm{MI}(Y,W) - \mathrm{MI}(Y,W|C), \mathrm{MI}(Y,X) - \mathrm{MI}(Y,X|C)\}$$
$$= -\max\{0-0, 0-(H(Y|C) - H(Y|X,C))\} = -\max\{0-0, 0-\ln(2)\}$$
$$= -\max\{0, -\ln(2)\} = 0.$$

This shows that, according to CMIM, both Y and Z can be selected at this step, whereas Y should be selected first, as confirmed by the target objective function values. The problem occurs because the class-relevant redundancy $\mathrm{MI}(Y,X|C)$ *brings a negative contribution to the term of the maximization that involves X, leading to* $\mathrm{TMI}(C,Y,X) = -ln(2)$*, thus forcing the maximum to be associated with the competing term, since* $\mathrm{TMI}(C,Y,W) = 0$*. As noted before, the maximum applied in this way penalizes the complementarity effects between Y and X that, as a result, are not reflected in the objective function of candidate Y; contrarily, the term that corresponds to an already selected feature that has no association with Y, i.e. the term involving W, is the one that is reflected in the objective function of candidate Y.*

Note that since $\mathrm{MI}(X_i, \boldsymbol{S}) \geq \mathrm{MI}(X_i, X_s)$ and $\mathrm{MI}(X_i, \boldsymbol{S}|C) \geq \mathrm{MI}(X_i, X_s|C)$, this approximation also undervalues both the inter-feature and the class-relevant redundancies. However, since the maximum is applied to the difference of the terms, it can no longer be concluded, as in the case of maxMIFS, that the approximation using a maximum is better than the one using an average (the case of JMI). In this case, the inter-feature redundancy term still pushes towards selecting the X_s that leads to the maximum value of $\mathrm{MI}(X_i, X_s)$, since it contributes positively to the value inside the maximum operator; contrarily, the class-relevant redundancy term pushes towards selecting X_s features that depart from the maximum value of $\mathrm{MI}(X_i, X_s|C)$, since it contributes negatively.

ACCEPTED MANUSCRIPT

JMIM JMIM uses the approximation based on the maximization operator, like maxMIFS and CMIM. However, the maximization embraces an additional term. Specifically,

$$\mathrm{TMI}(C, X_i, \boldsymbol{S}) - \mathrm{MI}(C, \boldsymbol{S}) \approx \max_{X_s \in \boldsymbol{S}} \{\mathrm{TMI}(C, X_i, X_s) - \mathrm{MI}(C, X_s)\}.$$

The additional term of JMIM, i.e. $\mathrm{MI}(C, X_s)$, tries to approximate a term of the target objective function that does not depend on X_i, i.e. $\mathrm{MI}(C, \boldsymbol{S})$, and brings additional problems to the selection process. We call this drawback *unimportant term approximated*. JMIM inherits the drawbacks of CMIM, complementarity penalized and redundancy undervalued. Moreover, the extra term adds a negative contribution to each X_s term, favoring X_s features with small association with C, which goes against the whole purpose of the feature selection process.

5.2. Properties of the methods

The drawbacks presented in Section 5.1 have consequences in terms of the good properties that forward feature selection methods must have, as expressed by Theorem 1 and Theorem 3: (i) the existence of meaningful bounds for the objective function, and (ii) the fact that fully relevant candidate features are the only ones that reach the maximum value of the objective function, while irrelevant and redundant features are the only ones to reach the minimum, which guarantees a perfect ordering of the features. With a few exceptions, the approximations taken by the various methods make them lose these properties. We will address these issues in the next two sections.

5.2.1. Preservation of bounds

We summarize the results relative to the preservation of the upper and lower bounds of the target objective function given by Theorem 1 in Table 5.

Methods that do not preserve the lower bound and the upper bound The methods MIFS, mRMR, maxMIFS, and CIFE, do not preserve neither the lower bound nor the upper bound. Indeed, the objective function of CIFE is unbounded, both superiorly and inferiorly, due to the overscaled redundancy

ACCEPTED MANUSCRIPT

drawback. Moreover, the objective functions of MIFS, mRMR, and maxMIFS are unbounded inferiorly, due to the lack of the compensation provided by the class-relevant redundancy (i.e. due to ignoring complementarity effects). For these methods, the upper bound of the objective function becomes $H(C)$. However, this bound is meaningless since it no longer expresses the uncertainty in C not explained by already selected features.

Table 5: Preservation of upper and lower bounds of the target objective function.

Bound	MIM	MIFS	mRMR	maxMIFS	CIFE	JMI	CMIM	JMIM
Upper-bound								X
Lower-bound	X					X	X	

Methods that preserve one bound MIM, JMI, CMIM, and JMIM preserve one of the bounds: JMIM preserves the upper bound and the remaining methods preserve the lower bound. MIM preserves the lower bound but just because its objective function is too simplistic.

In order to see that JMI preserves the lower bound, note that, using (17), its objective function can also be written as

$$\mathrm{OF}_{\mathrm{JMI}}(X_i) = \frac{1}{|\boldsymbol{S}|} \sum_{X_s \in \boldsymbol{S}} \mathrm{MI}(C, X_i | X_s). \tag{26}$$

For any candidate feature X_i, since $\mathrm{MI}(C, X_i|X_s) \geq 0$ for all $X_s \in \boldsymbol{S}$, it follows that $\mathrm{OF}_{\mathrm{JMI}}(X_i) \geq 0$.

For CMIM, using (21), it follows also from the non-negativity of $\mathrm{MI}(C, X_i|X_s)$ that $\mathrm{OF}_{\mathrm{CMIM}}(X_i) \geq 0$, again for any candidate feature X_i.

To see that JMIM preserves the upper bound, note that, using (22), its objective function can also be written as

$$\begin{aligned}
\mathrm{OF}_{\mathrm{JMIM}}(X_i) &= \min_{X_s \in \boldsymbol{S}} \{\mathrm{MI}(C, X_s) + \mathrm{MI}(C, X_i|X_s)\} \\
&= \min_{X_s \in \boldsymbol{S}} \{H(C) - H(C|X_i, X_s)\} \\
&= H(C) - \max_{X_s \in S} H(C|X_i, X_s).
\end{aligned}$$

ACCEPTED MANUSCRIPT

Hence, the objective function of JMIM has $H(C)$ as upper bound for any candidate feature X_i since $H(C|X_i, X_s) \geq 0$ for all $X_s \in \boldsymbol{S}$. Note that this is the desired bound since this method has target objective function OF as reference.

Despite maintaining one of the bounds stated by Theorem 1, MIM, JMI, CMIM, and JMIM, do not preserve the bound that involves the conditional entropy $H(C|\boldsymbol{S})$. For MIM the upper bound becomes $H(C)$ which is meaningless, as in the case of methods ignoring complementarity. For the remaining methods, the bound is lost due to the approximation that replaces the terms of the objective function that depend on set $\boldsymbol{S}$ by a function of the already selected features X_s taken individually. The lower bound of JMIM and the upper bounds of JMI and CMIM are now functions of $H(C|X_s)$. As in the case of MIFS, mRMR, and maxMIFS, the new bounds become meaningless: the upper bounds of JMI and CMIM no longer express the uncertainty in C that is not explained by the *complete* set of already selected features; and the lower bound of JMIM no longer expresses the uncertainty in C already explained by the *complete* set of already selected features.

In Section 6 we will illustrate the loss of bounds by the various methods.

5.2.2. Connection between bounds and feature types

Regarding the connections between the bounds of the objective functions and the feature types, stated by Theorem 3, these connections are lost for all methods, despite the fact that some bounds are preserved. It is no longer possible to assure that fully relevant features reach the maximum value of the objective function (when it exists) and that irrelevant and redundant features reach the minimum (when it exists). Moreover, the stopping criterion is lost. We will provide several examples in Section 6. This is again due to the approximation that replaces the dependencies on the whole set of already selected features, $\boldsymbol{S}$, by dependencies on individual features $X_s \in \boldsymbol{S}$, which is shared by all methods.

Redundant features given $\{X_s\}$ It would be useful to have results similar to those of Theorem 3, if their validity given X_s would imply their validity given $\boldsymbol{S}$. This is only meaningful for feature types such that if a feature has a

ACCEPTED MANUSCRIPT

type given X_s it will have the same type given $\boldsymbol{S}$. Unfortunately, this is only true for redundant features. In fact, according to Theorem 2, a feature that is redundant given X_s will also be redundant given $\boldsymbol{S}$. The same does not hold for irrelevant and relevant features since, as discussed in Section 4.2, as $\boldsymbol{S}$ grows, relevant features can become irrelevant, and vice-versa. Thus, only properties concerning redundancy given X_s are worth being considered. In this respect, a weaker version of Theorem 3 can be proved for CMIM.

Theorem 4. *If there exists $X_s \in \boldsymbol{S}$ such that X_i is a redundant feature given $\{X_s\}$, then $\mathrm{OF}_{\mathrm{CMIM}}(X_i) = 0$, i.e., the minimum possible value taken by the objective function of CMIM is reached.*

Proof. Since X_i is a redundant feature given $\{X_s\}$, then $\mathrm{MI}(C, X_i|X_s) = 0$ by (4) and (16). As a result, $\mathrm{OF}_{\mathrm{CMIM}}(X_i) = 0$ follows from (21). In fact, in order for $\min_{X_s \in \boldsymbol{S}} \mathrm{MI}(C, X_i|X_s)$ to be 0, it is enough that $\mathrm{MI}(C, X_i|X_s)$ is 0 for one particular X_s, since the terms involved in the minimization are all non-negative. □

Theorem 4 states that the objective function of CMIM reaches the minimum for a feature that is redundant given X_s. Note that a feature can be redundant given $\boldsymbol{S}$ but not redundant given X_s, which renders this result weaker than that of Theorem 3. Theorems analogous to Theorem 4 cannot be proved for the remaining methods, and we provide counter-examples in Section 6. The fact that only CMIM verifies this result is due to the redundancy undervalued drawback being less severe in this method.

To summarize, the approximations taken by all methods make them lose the good properties exhibited by the target objective functions, namely the guarantee that features are correctly ordered and the existence of a stopping criterion.

5.3. Other methods

We now briefly discuss other methods that have appeared in the literature, explaining why they have not been included as part of the representative meth-

ACCEPTED MANUSCRIPT

ods presented previously.

MIFS-U [39] differs from MIFS in the estimation of the MI between the candidate feature and the class—this is a meaningless difference for the type of theoretical properties of the methods that we intend to address, in which estimation does not play a role. MIFS-ND [40] considers the same reference terms as mRMR, employing a genetic algorithm to select the features, thus again not changing anything in theoretical terms. ICAP [41] is similar to CIFE, while forcing the terms $\mathrm{TMI}(C, X_i, X_s)$, $X_s \in \boldsymbol{S}$, to be seen as redundancy terms by only considering their contribution when they are positive (negative for the objective function). IGFS [42] chooses the same candidate features in each step as JMI; and CMIM-2 [43] is also just the same as JMI, as its objective function is defined exactly as (26).

A particular type of methods that were also not considered as representative is characterized by considering similar objective functions to those of the introduced representative methods, with the difference that all MI terms are replaced by corresponding normalized MI terms. More concretely: NMIFS [44] is an enhanced version of MIFS, MIFS-U, and mRMR; DISR [31] is adapted from JMI, and considers a type of normalization called symmetrical relevance; NJMIM [38] is adapted from JMIM, using also symmetrical relevance. Past experiments [cf. 18, 38] show that such normalizations make the methods more expensive, due to associated extra computations, with no compensation in terms of performance. In fact, it is argued in the mentioned experiments that the performance of such methods is actually worse than the performance of the corresponding methods that do not use normalized MI, which should be, as added by [18], related to the additional variance introduced by the estimation of the extra normalization term.

ACCEPTED MANUSCRIPT

6. Comparison of feature selection methods on a distributional setting

This section compares the feature selection methods using a distributional setting, based on a specific definition of class, features, and a performance metric. The setting provides an ordering for each of the methods, which is independent of specific datasets and estimation methods, and is compared with the ideal feature ordering. The aim of the setting is to illustrate how the drawbacks of the methods lead to incorrect feature ordering and to the loss of the good properties of the target objective functions.

We start by introducing a performance measure for feature selection methods that does not rely on the specificities of a fixed classifier—the *minimum Bayes risk* (MBR). We then describe the characteristics of the setting, namely the definitions of class and features, and show how the quantities required to calculate the objective functions of the methods, i.e. the various types of MI, are calculated. Finally, we present and discuss the results.

6.1. Minimum Bayes risk

Commonly, the performance measures used to compare forward selection methods depend on how a particular classifier performs for certain data sets. As a result, it is not clear if the obtained conclusions are exclusively explained by the characteristics of the feature selection method, or if the specificities of the classifier and/or the data under study create confounding effects. To overcome this limitation, we consider a different type of performance measure that is computed at each step of the forward selection method under consideration. Using the set of selected features until a given step, we obtain, for a fixed classifier, the associated *Bayes risk* (BR) or *total probability of misclassification* [45, Ch. 11]. Bayes risk is a theoretical measure in the sense that it does not rely on data but instead directly on the, assumed to be known, distributions of the involved features into consideration [46, 47, for practical contexts where it was used]. The *Bayes classifier* [22, Ch. 1] is a classifier that defines a classification

ACCEPTED MANUSCRIPT

rule associated with the minimum Bayes risk, which will be our performance measure. The suitability of this measure to our setting results from the fact that it relies on the distributions of the features, and also on their class-conditional distributions.

6.1.1. Bayes risk and Bayes classifier

For a given class C, with values on the set $\{0, 1, ..., c\}$, and a set of selected features $\boldsymbol{S}$, with support $\mathcal{S}$, a $(C, \boldsymbol{S})$-classifier g is a (Borel-)measurable function from $\mathcal{S}$ to $\{0, 1, ..., c\}$, and $g(\boldsymbol{s})$ denotes the value of C to which the observation $\boldsymbol{s}$ is assigned by the classifier. Then, the Bayes risk of the $(C, \boldsymbol{S})$-classifier g is given by

$$\mathrm{BR}(C, \boldsymbol{S}, g) = P(g(\boldsymbol{S}) \neq C)) = \sum_{j=0}^{c} P(g(\boldsymbol{S}) \neq j | C = j) P(C = j).$$

The proposed performance evaluation measure consists of the minimum possible value of the BR, called *minimum Bayes risk* (MBR). Thus, for a given class C and a set of selected features $\boldsymbol{S}$, the associated minimum Bayes risk, $\mathrm{MBR}(C, \boldsymbol{S})$, is given by:

$$\mathrm{MBR}(C, \boldsymbol{S}) = \min_{g} \mathrm{BR}(C, \boldsymbol{S}, g).$$

The minimum Bayes risk corresponds to the Bayes risk of the so-called *Bayes classifier*. The $(C, \boldsymbol{S})$ Bayes classifier assigns an object $\boldsymbol{s} \in \mathcal{S}$ to the value that C is most likely to take given that $\boldsymbol{S} = \boldsymbol{s}$. That is, the $(C, \boldsymbol{S})$ Bayes classifier g is such that

$$\begin{aligned} g(\boldsymbol{s}) &= \underset{j \in \{0,1,...,c\}}{\operatorname{argmax}} P(C = j | \boldsymbol{S} = \boldsymbol{s}) \\ &= \underset{j \in \{0,1,...,c\}}{\operatorname{argmax}} P(C = j) f_{\boldsymbol{S}|C=j}(\boldsymbol{s}). \end{aligned}$$

Note that, in particular, when there are two possible values for the class, i.e. $c = 1$, the Bayes classifier g is such that [45, Ch. 11]:

$$g(\boldsymbol{s}) = 1 \iff \frac{f_{\boldsymbol{S}|C=0}(\boldsymbol{s})}{f_{\boldsymbol{S}|C=1}(\boldsymbol{s})} \leq \frac{P(C = 1)}{P(C = 0)}. \tag{27}$$

ACCEPTED MANUSCRIPT

6.1.2. Properties of the minimum Bayes risk

We now discuss a few properties of the minimum Bayes risk, the proposed performance evaluation criterion. In the following, measurable should be read as Borel-measurable.

Theorem 5. *If C is a measurable function of $\boldsymbol{S}$, then* $\mathrm{MBR}(C, \boldsymbol{S}) = 0$.

Proof. Let g be the measurable function such that $C = g(\boldsymbol{S})$. As $C = g(\boldsymbol{S})$, it follows that $\mathrm{BR}(C, \boldsymbol{S}, g) = P(g(\boldsymbol{S}) \neq C)) = 0$. As $\mathrm{MBR}(\mathrm{C}, \boldsymbol{S})$ is non-negative and $\mathrm{MBR}(\mathrm{C}, \boldsymbol{S}) \leq \mathrm{BR}(\mathrm{C}, \boldsymbol{S}, \mathrm{g})$, we conclude that $\mathrm{MBR}(C, \boldsymbol{S}) = 0$, as intended. □

Note that C being a measurable function of $\boldsymbol{S}$ is equivalent to saying that features in $\boldsymbol{S}$ fully explain the class.

Theorem 6. *If X_i is a measurable function of $\boldsymbol{S}$, then* $\mathrm{MBR}(C, \boldsymbol{S} \cup \{X_i\}) = \mathrm{MBR}(C, \boldsymbol{S})$.

Proof. Let ξ be the measurable function such that $X_i = \xi(\boldsymbol{S})$, and g be the $(C, \boldsymbol{S} \cup \{X_i\})$ Bayes classifier, so that, in particular, $\mathrm{MBR}(C, \boldsymbol{S} \cup \{X_i\}) = \mathrm{BR}(C, \boldsymbol{S} \cup \{X_i\}, g)$. Let g' be the $(C, \boldsymbol{S})$ classifier such that, given an observation $\boldsymbol{s} \in \mathcal{S}$, $g'(\boldsymbol{s}) = j$ when $g(\boldsymbol{s}, \xi(\boldsymbol{s})) = j$. Then $\mathrm{BR}(C, \boldsymbol{S} \cup \{X_i\}, g) = \mathrm{BR}(C, \boldsymbol{S}, g')$. As a consequence, by transitivity, $\mathrm{MBR}(C, \boldsymbol{S} \cup \{X_i\}) = \mathrm{BR}(C, \boldsymbol{S}, g')$. This implies that $\mathrm{MBR}(C, \boldsymbol{S}) \leq \mathrm{MBR}(C, \boldsymbol{S} \cup \{X_i\})$. In turn, it always holds that $\mathrm{MBR}(C, \boldsymbol{S}) \geq \mathrm{MBR}(C, \boldsymbol{S} \cup \{X_i\})$ since $\boldsymbol{S} \subset \boldsymbol{S} \cup \{X_i\}$. Therefore, $\mathrm{MBR}(C, \boldsymbol{S} \cup \{X_i\}) = \mathrm{MBR}(C, \boldsymbol{S})$. □

Note that X_i being a measurable function of $\boldsymbol{S}$ is equivalent to saying that X_i is redundant given $\boldsymbol{S}$.

6.2. Setting description

We now describe the distributional setting used to illustrate the various deficiencies of the feature selection methods. The class chosen for this setting

ACCEPTED MANUSCRIPT

is a generalization of the one proposed by [27], which was based on the scenario introduced by [39] and later used by [22]. It is defined as

$$C_k = \begin{cases} 0, & X + kY < 0 \\ 1, & X + kY \geq 0 \end{cases}, \tag{28}$$

where X and Y are independent features with standard normal distributions and $k \in]0, +\infty[$.

According to the discussion in Section 4, our scenario includes fully relevant, relevant, redundant, and irrelevant features. Specifically, our features are X, $X - k'Y$, $k' > 0$, Z and X_{disc}. X and $X - k'Y$ were chosen as relevant features that, taken together, fully explain the class. As irrelevant feature, we chose Z, independent of X and Y, which for simplicity is considered to follow a Bernoulli distribution with success probability $1/2$. Finally, as redundant feature we chose

$$X_{\text{disc}} = \begin{cases} 0, & X < 0 \\ 1, & X \geq 0 \end{cases}, \tag{29}$$

The first selected feature is the candidate X_i that has the largest value of $\text{MI}(C_k, X_i)$. The possible candidates are X, $X - k'Y$, and X_{disc}, which are the initially relevant features. Z is an irrelevant feature and, therefore, will not be selected first. We want X to be selected first to assure that, at the second step of the algorithm, there will be, as candidates, one fully relevant, one redundant, and one irrelevant feature. This provides an extreme scenario, where the relevancy level of the relevant feature is the maximum possible, making a wrong decision the hardest to occur. We next discuss the conditions for selecting X before $X - k'Y$ and before X_{disc}.

X is selected before $X - k'Y$ if $\text{MI}(C_k, X) > \text{MI}(C_k, X - k'Y)$, which is equivalent to the condition

$$\arctan k < (\pi - \arctan k') - \arctan k,$$

where the left term represents the angle between the lines $X = 0$ and $X + kY = 0$ and the right term represents the angle between the lines $X - k'Y = 0$ and

ACCEPTED MANUSCRIPT

$X + kY = 0$, in the context of the two-dimensional space defined by the pair of orthogonal vectors (X, Y). The condition can be written in terms of k as

$$k < \tan\left(\frac{\pi - \arctan k'}{2}\right). \tag{30}$$

Feature X_{disc} is never selected before X since $\text{MI}(C_k, X_{\text{disc}}) \leq \text{MI}(C_k, X)$ for all $k > 0$. To see this, note that this inequality can be written, using (4), as $H(C_k) - H(C_k|X_{\text{disc}}) \leq H(C_k) - H(C_k|X)$, which is equivalent to $H(C_k|X_{\text{disc}}) \geq H(C_k|X)$. This is equivalent to $H(C_k|X_{\text{disc}}) \geq H(C_k|X_{\text{disc}}, X)$, which holds by (16). In turn, $H(C_k|X) = H(C_k|X_{\text{disc}}, X)$ is equivalent to $\text{MI}(C_k, X_{\text{disc}}|X) = 0$ by (12). Finally, since X_{disc} is redundant given $\{X\}$, equations (12) and (16) can be used to verify that $\text{MI}(C_k, X_{\text{disc}}|X) = 0$.

In view of the above discussion, the ideal feature ordering coming out of the distributional setting is X in first place and $X - k'Y$ in second place. Ideally, the feature selection method should stop at this step, since a fully relevant feature has been found. However, further steps need to be considered, since actual methods do not preserve the stopping criterion, as discussed in Section 5.2. Then, at the third step, the remaining features, Z and X_{disc}, must be equally likely to be selected since, according to Theorems 3.2 and 3.3, both their target objective functions reach the lower bound.

6.3. Required quantities

In order to be able to determine the order in which the features are selected by the different methods, we have to derive expressions, depending on k and k', needed for evaluating the corresponding objective functions. We need: the MI between each candidate feature and the class, the MI between different pairs of candidate features, and the class-conditional MI between pairs of candidate features. The computation of these quantities require obtaining the univariate entropies of the candidate features and of the class. The derivations of such expressions are provided in Appendix Appendix A.1 and their final forms are available in Tables 6 to 9.

ACCEPTED MANUSCRIPT

Table 6: Entropies of the class, C_k, and the input features.

	C_k	X	$X - k'Y$	Z	X_{disc}
Entropy	$\ln(2)$	$\frac{1}{2}\ln(2\pi e)$	$\frac{1}{2}\ln(2\pi e(1+k'^2))$	$\ln(2)$	$\ln(2)$

Table 7: MI between each input feature and the class, C_k.

A	$\text{MI}(C_k, A)$
X	$\frac{1}{2}\ln(2\pi e) - \frac{1}{2}\sum_{j=0}^{1}\int_{\mathbb{R}} f_{X\|C_k=j}(u)\ln f_{X\|C_k=j}(u)du$
$X - k'Y$	$\frac{1}{2}\ln(2\pi e(1+k'^2)) - \frac{1}{2}\sum_{j=0}^{1}\int_{\mathbb{R}} f_{X-k'Y\|C_k=j}(u)\ln f_{X-k'Y\|C_k=j}(u)du$ [a]
Z	0
X_{disc}	$2\ln(2) + \frac{\arctan k}{\pi}\ln(\frac{\arctan k}{2\pi}) + (1 - \frac{\arctan k}{\pi})\ln(\frac{1}{2} - \frac{\arctan k}{2\pi})$ [b]

[a] $X|C_k = j \sim \text{SN}(0, 1, \frac{(-1)^{j+1}}{k})$, $j = 0, 1$.
[b] $X - k'Y|C_k = j \sim \text{SN}(0, \sqrt{1+k'^2}, (-1)^{j+1}(\frac{1-kk'}{k+k'}))$, $j = 0, 1$.

Univariate entropies.. We start with a summary of the univariate entropies of the different features and of the class presented in Table 6. The corresponding derivations can be found in Appendix Appendix A.1.

MI between input features and the class.. As for the MI between input features and the class, they are provided in Table 7. The corresponding derivations can be found in Appendix Appendix A.2.

It must be added that the notation $W \sim \text{SN}(\mu, \sigma, \alpha)$ (where $\mu \in \mathbb{R}$, $\sigma > 0$, and $\alpha \in \mathbb{R}$), means that the random variable W follows a skew-normal distribution, so that it has probability density function [48]

$$f_W(w) = \frac{2}{\sigma}\phi(\frac{w-\mu}{\sigma})\Phi(\frac{\alpha(w-\mu)}{\sigma}), \quad w \in \mathbb{R}, \tag{31}$$

where $\Phi(z)$ denotes the value of the standard normal distribution function at point z, while $\phi(z)$ denotes the probability density function, for the same distribution, also at z.

MI between pairs of input features.. As for the MI between the different pairs of input features, they are provided in Table 8. The corresponding derivations

ACCEPTED MANUSCRIPT

Table 8: MI between pairs of input features.

A	B	$\mathrm{MI}(\cdot,\cdot)$
X	$X - k'Y$	$\frac{1}{2}\ln(1+\frac{1}{k'^2})$
X	X_{disc}	$\ln(2)$
$X - k'Y$	X_{disc}	$\frac{1}{2}\ln(2\pi e) - \frac{1}{2}\sum_{j=0}^{1}\int_{\mathbb{R}} f_{X\|C_{k'}=j}(u)\ln f_{X\|C_{k'}=j}(u)du$ [a]
Z	B	$0, \quad B \in \{X, X-k'Y, X_{\mathrm{disc}}\}$

[a] $X|C_{k'}=j \sim \mathrm{SN}(0,1,\frac{(-1)^{j+1}}{k'})$, $j=0,1$.

Table 9: Class-conditional MI between pairs of input features.

A	B	$\mathrm{MI}(\cdot,\cdot\|C_k)$
X	$X - k'Y$	$\frac{1}{2}\sum_{j=0}^{1}\int_{\mathbb{R}} f_{X\|C_k=j}(u)\ln f_{X\|C_k=j}(u)du +$ $\frac{1}{2}\sum_{j=0}^{1}\int_{\mathbb{R}} f_{X-k'Y\|C_k=j}(u)\ln f_{X-k'Y\|C_k=j}(u)du -$ $(1+\ln\pi+\ln k')$ [a,b]
X	X_{disc}	$-\frac{\arctan k}{\pi}\ln(\frac{\arctan k}{\pi}) - (1-\frac{\arctan k}{\pi})\ln(1-\frac{\arctan k}{\pi})$
$X - k'Y$	X_{disc}	$\frac{1}{2}\sum_{j=0}^{1}\int_{\mathbb{R}} f_{X-k'Y\|C_k=j}(u)\ln f_{X-k'Y\|C_k=j}(u)du - h(X-k'Y\|X_{\mathrm{disc}},C_k)$ [b,c]
Z	B	$0, \quad B \in \{X, X-k'Y, X_{\mathrm{disc}}\}$

[a] $X|C_{k'}=j \sim \mathrm{SN}(0,1,\frac{(-1)^{j+1}}{k'})$, $j=0,1$. [b] $X-k'Y|C_k=j \sim \mathrm{SN}(0,\sqrt{1+k'^2},(-1)^{j+1}(\frac{1-kk'}{k+k'}))$, $j=0,1$. [c] $h(X-k'Y|X_{\mathrm{disc}},C_k)$ in (A.5).

765 can be found in Appendix Appendix A.3.

Class-conditional MI between pairs of input features.. As for the class-conditional MI between the different pairs of input features, they are provided in Table 9. The corresponding derivations can be found in Appendix Appendix A.4.

6.4. Applying the different feature selection methods

We now present the results of applying the various feature selection methods to the distributional setting. The feature ordering will be discussed for different values of k and k'. Taking the objectives of this study into consideration, for

ACCEPTED MANUSCRIPT

fixed k', the most interesting case is that where X alone leaves the largest possible amount of information undetermined about the class; this leads to $X - k'Y$ having the most importance in the explanation of the class, making the error of not choosing it after X the worst possible. According to the performance metric introduced in Section 6.1, we want to choose a value of k that leads to a large MBR when X is the only selected feature, $\mathrm{MBR}(C_k, \{X\})$, which is given by (see Appendix Appendix B):

$$\mathrm{MBR}(C_k, \{X\}) = \frac{\arctan k}{\pi}. \tag{32}$$

Since $\mathrm{MBR}(C_k, \{X\})$ is an increasing function of k, we want k to be as large as possible, under the restriction (30). We consider $k = \tan\big((\pi - \arctan k' - 10^{-6})/2\big)$.

Given that, in our setting, the features X and $X - k'Y$ fully explain the class, and that, according to Theorem 5, $\mathrm{MBR}(C_k, \{X, X - k'Y\}) = 0$, it makes sense to take the MBR based on the first two selected features as performance measure for characterizing each forward feature selection method. This measure is denoted by MBR_2.

We will carry out two different studies. In the first one, we concentrate on the methods that ignore complementarity; i.e. MIFS, mRMR, and maxMIFS, and study the feature ordering as a function of k'. The purpose of this study is to highlight the consequences of ignoring complementarity. In the second study, we compare the feature ordering of all methods under analysis, for fixed k and k'. The goal is to provide examples showing wrong decisions made by the various methods.

6.4.1. The consequences of ignoring complementarity

In this section, we address the consequences of ignoring complementarity, as a function of k'; thus, we concentrate on methods MIFS ($\beta = 1$), mRMR, and maxMIFS. The motivation for scanning k' is that it provides different levels of association between the already selected feature X and the candidate feature $X - k'Y$. For small values of k', X and $X - k'Y$ are strongly associated, and the level of association decreases as k' increases.

ACCEPTED MANUSCRIPT

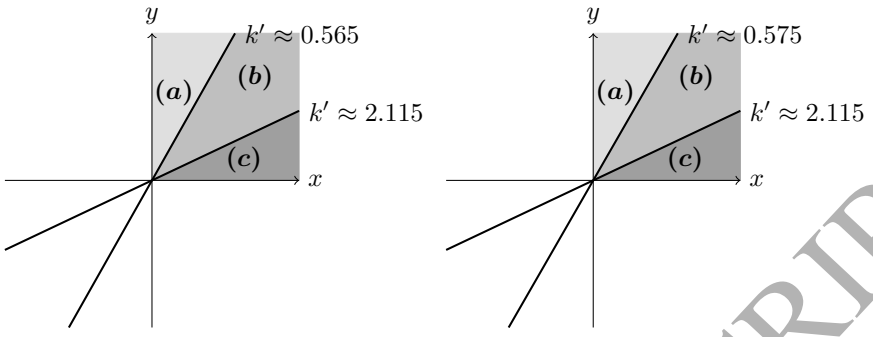


Figure 1: Regions associated with a specific ordering of the features, defined by the values of k'. In this representation, k' is defined through the line $x - k'y = 0$. For region (a), the ordering is $\{X, Z, X_{\text{disc}}, X - k'Y\}$; for (b), it is $\{X, Z, X - k'Y, X_{\text{disc}}\}$; and for (c), the ordering is already correct since X is chosen first and $X - k'Y$ second. For methods MIFS and maxMIFS (resp. mRMR), represented in the left (resp. right), (a) is associated with $0 < k' < 0.575$ (resp. $0 < k' < 0.565$) and (b) with $0.575 < k' < 2.115$ (resp. $0.565 < k' < 2.115$). Region (c) is associated with $k' > 2.115$ for the three methods.

Scanning k' from 0 to $+\infty$ defines three regions, each corresponding to a specific feature ordering. This is shown in Figure 1, where k' was scanned with a step size of 0.01, starting at 0.01. To complete the discussion, we include the objective function values in the second step of the algorithms in Figure 2, and the corresponding MBR_2 values in Figure 3. Note that, at this step, the objective function takes the same value for all candidate features and methods and MBR_2 takes the same value for all methods.

For small values of k', smaller than 0.565 for MIFS and maxMIFS, and than 0.575 for mRMR—region (a)—the feature ordering is X, Z, X_{disc}, $X - k'Y$. In this region, $X - k'Y$ is chosen last due to a large inter-feature redundancy with X. As shown in Figure 2, in this region, the objective function values of $X - k'Y$ and X_{disc} at the second step are negative and smaller than that of Z, which explains why Z is selected in second place. At the third step, the objective function values of $X - k'Y$ and X_{disc} are exactly the same as in the second step for MIFS and maxMIFS, and only slightly different for mRMR. Thus, in this region, the objective functions of $X - k'Y$ are more negative than those of X_{disc},

ACCEPTED MANUSCRIPT

which explains why X_{disc} is selected in third place.

For intermediate values of k', smaller than 2.115, and larger than 0.565 for MIFS and maxMIFS and than 0.575 for mRMR—region (b)—the feature ordering is X, Z, $X - k'Y$, X_{disc}. In this region, the objective functions of $X - k'Y$ are larger than those of X_{disc}, but smaller than those of Z.

For large values of k', larger than 2.115—region (c)—the correct feature ordering is achieved since $X - k'Y$ is selected in second place. Note that in this region, there are two possible orderings for Z and X_{disc}, but this issue is not relevant for our discussion.

The problem of these methods in regions (a) and (b) is due to the lack of the class-relevant redundancy term in their objective functions, which expresses the complementarity effects. In fact, the association between X and $X - k'Y$, as measured by $\text{MI}(X - k'Y, X)$, grows significantly as k' approaches 0, but so does the class-relevant redundancy, which is given by $\text{MI}(X - k'Y, X|C_k)$. Ignoring the compensation given by the latter term leads to objective function values that can take negative values. Moreover, this explains why the lower bound of 0 from Theorem 1, associated with the target objective function, is lost for these methods. Also, in contradiction with the good properties of the target objective function, the objective functions of these methods do not take the same (minimum) values at X_{disc} and Z. The MBR_2 values of these methods (see Figure 3) confirm that the performance is very poor in regions (a) and (b): it is above 0.4 in region (a) and above 0.3 in region (b). These results show that ignoring complementarity is a severe drawback that can lead to gross errors in the feature selection process.

Figure 2 also shows that results analogous of Theorem 4 do not hold for MIFS, mRMR, and maxMIFS. In fact, the objective function at X_{disc}, a redundant feature, is not necessarily the minimum; in particular, this happens for small values of k', where the objective function at $X - k'Y$ takes lower values than that at X_{disc}.

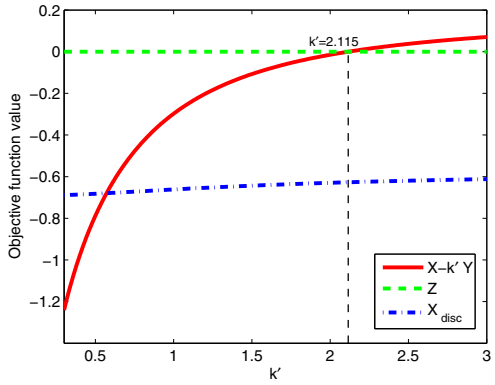


Figure 2: Evaluation of the objective function for the different candidate features in the second step of the algorithms (MIFS, mRMR, and maxMIFS) depending on the value of k'.

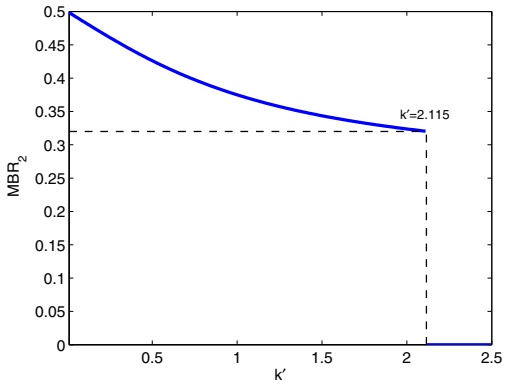


Figure 3: MBR_2 for the different algorithms (MIFS, mRMR, and maxMIFS) depending on the value of k'. For $k' < 2.115$, $\mathrm{MBR}_2 = \frac{\pi - \arctan k' - 10^{-6}}{2\pi}$ using $k = \tan\left((\pi - \arctan k' - 10^{-6})/2\right)$ and (32); for $k' > 2.115$, the right second feature is chosen, $X - k'Y$, so that $\mathrm{MBR}_2 = 0$.

6.4.2. Feature ordering

We now compare the feature ordering of all methods, for fixed k and k'. In order to place the methods in a challenging situation, we use k and k' values that maximize the MBR. Per (32) we need to maximize k and per (30) we need to minimize k'. We choose for k' the first value of the grid used in the

ACCEPTED MANUSCRIPT

Table 10: Feature ordering and corresponding MBR_2, for $k = 199.985$ and $k' = 0.01$.

Methods	Order of feature selection				MBR_2
MIM	X	$X - k'Y$	X_{disc}	Z	0
MIFS ($\beta = 1$)	X	Z	X_{disc}	$X - k'Y$	0.498
mRMR	X	Z	X_{disc}	$X - k'Y$	0.498
maxMIFS	X	Z	X_{disc}	$X - k'Y$	0.498
CIFE	X	$X - k'Y$	X_{disc}	Z	0
JMI	X	$X - k'Y$	X_{disc}	Z	0
CMIM	X	$X - k'Y$	Z/X_{disc}	X_{disc}/Z	0
JMIM	X	$X - k'Y$	X_{disc}	Z	0

Table 11: MI between the class and each input feature, for $k = 199.985$ and $k' = 0.01$.

	X	$X - k'Y$	Z	X_{disc}
$\mathrm{MI}(\cdot, C_k)$	≈ 0	≈ 0	0	≈ 0

context of Figure 1, i.e. $k' = 0.01$. In this case, $k = \tan(\frac{\pi - \arctan k' - 10^{-6}}{2}) = 199.985$ and $\mathrm{MBR}(C_k, \{X\}) \approx 0.498$. Recall that, since (30) holds, and therefore $\mathrm{MI}(C_k, X) > \mathrm{MI}(C_k, X - k'Y)$, X is always selected in first place.

Table 10 shows the feature ordering and the associated values of MBR_2. The ordering of features relates to the concrete values of the terms that compose the objective functions. These are provided in Table 11, which contains the values of MI between each candidate feature and the class; Table 12, which contains the values of MI between the different features; and Table 13, which contains the values of the class-conditional MI between the different input features. Note that, in Table 11, $\mathrm{MI}(C_k, X)$, $\mathrm{MI}(C_k, X - k'Y)$, and $\mathrm{MI}(C_k, X_{\mathrm{disc}})$ are all shown as taking approximately the value 0, but actually $\mathrm{MI}(C_k, X)$ is the largest one.

Table 10 shows that all methods, except MIFS, mRMR, and maxMIFS, achieve an MBR_2 of 0. However, the third step of the algorithm is only completely correct for CMIM. In fact, it should be equally likely to choose Z or X_{disc}, but CIFE, JMI, JMIM, and MIM select X_{disc} first.

ACCEPTED MANUSCRIPT

Table 12: MI between pairs of input features, for $k = 199.985$ and $k' = 0.01$.

$\mathrm{MI}(\cdot,\cdot)$	X	$X - k'Y$	Z
$X - k'Y$	4.605		
Z	0	0	
X_{disc}	0.693	0.686	0

Table 13: Class-conditional MI between pairs of input features, for $k = 199.985$ and $k' = 0.01$.

$\mathrm{MI}(\cdot,\cdot\|C_k)$	X	$X - k'Y$	Z
$X - k'Y$	5.298		
Z	0	0	
X_{disc}	0.693	0.689	0

MIM suffers from redundancy ignored drawback. The fact that the selection is correct at the first two steps of the feature selection process is meaningless; it only happens because $\mathrm{MI}(C_k, X - k'Y)$ is slightly larger than $\mathrm{MI}(C_k, X_{\mathrm{disc}})$.

The methods that ignore complementarity, i.e. MIFS, mRMR, and maxMIFS, fail at the second step of the feature selection process, by not selecting $X - k'Y$. For all methods, the objective function is 0 for Z, $\mathrm{MI}(C_k, X - k'Y) - \mathrm{MI}(X - k'Y, X) = -4.605$ for $X - k'Y$, and $\mathrm{MI}(C_k, X_{\mathrm{disc}}) - \mathrm{MI}(X_{\mathrm{disc}}, X) = -0.693$ for X_{disc}, which explains why Z is selected at this step. Adding the class-relevant redundancy term to the objective functions, would make them take the value $\ln(2)$ for $X - k'Y$ and 0 for X_{disc}, leading to the selection of $X - k'Y$. In fact, the class-relevant redundancy term is $\mathrm{MI}(X - k'Y, X|C_k) = 5.298$ for $X - k'Y$, and $\mathrm{MI}(X_{\mathrm{disc}}, X|C_k) = 0.693$ for X_{disc}. Note that $\ln(2)$ is precisely the maximum of the target objective function, which is achieved for fully relevant features (the case of $X - k'Y$), and the minimum is 0, achieved by irrelevant and redundant features (the cases of Z and X_{disc}). Thus, accounting for the class-relevant redundancy compensates the potentially large negative values associated with the inter-feature redundancy.

With the exception of CMIM, the methods that do not ignore complemen-

ACCEPTED MANUSCRIPT

tarity, fail at the third step of the feature selection process by preferring X_{disc} over Z, as shown in Table 10.

As discussed in Section 5, CIFE suffers from overscaled redundancy drawback. At the third step of the feature selection process, after selecting X and $X - k'Y$, the objective function for candidate feature X_i is

$$\text{MI}(C_k, X_i) - \text{MI}(X_i, X) + \text{MI}(X_i, X|C_k) - \text{MI}(X_i, X - k'Y) + \text{MI}(X_i, X - k'Y|C_k), \tag{33}$$

while the associated target objective function is

$$\text{MI}(C_k, X_i) - \text{MI}(X_i, \{X, X - k'Y\}) + \text{MI}(X_i, \{X, X - k'Y\}|C_k). \tag{34}$$

Both objective functions take the value 0 for the candidate feature Z. For the candidate X_{disc}, the target objective function (34) can be written as

$$\text{MI}(C_k, X_i) - \text{MI}(X_i, X) + \text{MI}(X_i, X|C_k), \tag{35}$$

given that $\text{MI}(X_{\text{disc}}, \{X, X - k'Y\}) = \text{MI}(X_{\text{disc}}, X)$ and $\text{MI}(X_{\text{disc}}, \{X, X - k'Y\}|C_k) = \text{MI}(X_{\text{disc}}, X|C_k)$. Concerning the first condition, note that $\text{MI}(X_i, \boldsymbol{S}) = \text{MI}(X_{\text{disc}}, \{X, X - k'Y\}) = H(X_{\text{disc}}) - H(X_{\text{disc}}|X, X - k'Y) = H(X_{\text{disc}}) - H(X_{\text{disc}}|X) = \text{MI}(X_{\text{disc}}, X)$, since $H(X_{\text{disc}}|X) = 0$ implies that $H(X_{\text{disc}}|X, X - k'Y) = 0$ also, by (16); a similar reasoning can be used to show that the second condition also holds.

Thus, in the case of X_{disc}, we see that, when comparing the objective function of CIFE, given by (33), with the target objective function, given by (35), CIFE includes an extra part with two terms, $-\text{MI}(X_i, X - k'Y) + \text{MI}(X_i, X - k'Y|C_k)$, which is responsible for increasing the redundancy scale. In our case, the extra term takes the value $-\text{MI}(X_{\text{disc}}, X - k'Y) + \text{MI}(X_{\text{disc}}, X - k'Y|C_k) = -0.686 + 0.689 = 0.003$; recall Tables 12 and 13. This is exactly the value of the objective function at X_{disc}, since the remaining terms sum to 0, which explains why X_{disc} is selected before Z. To see that the remaining terms sum to 0, note that these terms correspond exactly to the evaluation of the target objective function OF', and recall that the target objective function value must be 0 for a redundant feature.

ACCEPTED MANUSCRIPT

The amount of overscaling is relatively modest in this example but, clearly, the problem gets worse as $\boldsymbol{S}$ increases, since more terms are added to the objective function. We also note that, while in this case the objective function has been overvalued, it could have equally been undervalued. This fact together with the overscaling problem is what makes the objective function of CIFE not bounded, neither from below nor from above.

JMI tried to overcome the problem of CIFE by introducing the scaling factor $1/|\boldsymbol{S}|$ in the TMI approximation. However, as discussed in Section 5, this leads to redundancy undervalued drawback. At the third step of the feature selection process, the objective function of JMI is

$$\mathrm{MI}(C_k, X_i) - \frac{1}{2}\mathrm{MI}(X_i, X) + \frac{1}{2}\mathrm{MI}(X_i, X|C_k) - \frac{1}{2}\mathrm{MI}(X_i, X - k'Y) + \frac{1}{2}\mathrm{MI}(X_i, X - k'Y|C_k)$$

for the candidate X_i. Its value equals 0 for the candidate Z, but for candidate X_{disc} it equals $0 - 0.5 \times 0.693 + 0.5 \times 0.693 - 0.5 \times 0.689 + 0.5 \times 0.686 = 0.0015$, which explains why X_{disc} is selected before Z.

This results directly from the undervaluing of the terms $\mathrm{MI}(X_i, \boldsymbol{S})$ and $\mathrm{MI}(X_i, \boldsymbol{S}|C)$ of the target objective function at $X_i = X_{\mathrm{disc}}$. In fact, $\mathrm{MI}(X_i, \boldsymbol{S}) = \mathrm{MI}(X_{\mathrm{disc}}, X) = 0.693$, but JMI approximates it by a smaller value, i.e. $\frac{1}{2}\mathrm{MI}(X_{\mathrm{disc}}, X) + \frac{1}{2}\mathrm{MI}(X_{\mathrm{disc}}, X - k'Y) = \frac{1}{2} \times 0.693 + \frac{1}{2} \times 0.686 = 0.6895$. Similarly, $\mathrm{MI}(X_{\mathrm{disc}}, \boldsymbol{S}|C_k) = \mathrm{MI}(X_{\mathrm{disc}}, X|C_k) = 0.693$, but again JMI approximates it by a smaller value, i.e. $\frac{1}{2}\mathrm{MI}(X_{\mathrm{disc}}, X|C_k) + \frac{1}{2}\mathrm{MI}(X_{\mathrm{disc}}, X - k'Y|C_k) = \frac{1}{2} \times 0.693 + \frac{1}{2} \times 0.689 = 0.691$.

JMIM introduced an additional term in the objective function which, as discussed in Section 5, is unimportant and may lead to confusion in the selection process—unimportant term approximated drawback. At the third step of the selection process, the objective function of JMIM is

$$\begin{aligned}\mathrm{MI}(C_k, X_i) - \max\{&\mathrm{MI}(X_i, X) - \mathrm{MI}(X_i, X|C_k) - \mathrm{MI}(C_k, X), \\ &\mathrm{MI}(X_i, X - k'Y) - \mathrm{MI}(X_i, X - k'Y|C_k) - \mathrm{MI}(C_k, X - k'Y)\},\end{aligned}$$

for candidate feature X_i. In this case, the objective function for candidate feature Z is $0 - \max\{0 - 0 - \mathrm{MI}(C_k, X), 0 - 0 - \mathrm{MI}(C_k, X - k'Y)\}$, and for candidate X_{disc} it is $0 - \max\{0.693 - 0.693 - \mathrm{MI}(C_k, X), 0.686 - 0.689 - \mathrm{MI}(C_k, X - k'Y)\}$.

ACCEPTED MANUSCRIPT

We first note that $\mathrm{MI}(C_k, X)$ and $\mathrm{MI}(C_k, X - k'Y)$ are both approximately 0, while $\mathrm{MI}(C_k, \{X, X - k'Y\})$, the quantity they try to approximate, takes the value $\ln(2)$. Since, per design of our experiment $\mathrm{MI}(C_k, X) > \mathrm{MI}(C_k, X - k'Y)$, it turns out that the objective function of Z equals $\mathrm{MI}(C_k, X - k'Y)$, and that of X_{disc} equals $\mathrm{MI}(C_k, X)$, leading to the selection of X_{disc}. There are two observations that should be pointed out. First, contrarily to the previous cases of CIFE and JMI, the objective function for Z takes a value that is no longer according to the corresponding target objective function, which in this case should be $\mathrm{MI}(C_k, \{X, X - k'Y\}) = \ln(2)$. Second, the choice between the two features, Z and X_{disc}, is being done by two terms, $\mathrm{MI}(C_k, X)$ and $\mathrm{MI}(C_k, X - k'Y)$, that try to approximate a term that does not depend on the candidate features, $\mathrm{MI}(C, \boldsymbol{S})$, and therefore should take the same value for both features and not become a deciding factor.

The results regarding MIM, CIFE, JMI, and JMIM provide counter-examples showing that theorems analogous to Theorem 4 do not hold for these methods. Indeed, in all cases, the objective function at X_{disc}, a redundant feature, takes values different from the minimum of the corresponding objective function.

CMIM is the only method that performs correctly in the distributional setting. At the third step of the feature selection process, the objective function is 0 for both Z and X_{disc}. The latter result can be obtained from Theorem 4, since X_{disc} is redundant given $\{X\}$. This can be confirmed numerically. The objective function of CMIM at the third step of the feature selection process is

$$\mathrm{MI}(C_k, X_i) - \max\left\{\mathrm{MI}(X_i, X) - \mathrm{MI}(X_i, X|C_k), \mathrm{MI}(X_i, X - k'Y) - \mathrm{MI}(X_i, X - k'Y|C_k)\right\},$$

for candidate feature X_i. In this case, the objective function for candidate feature Z is 0, and the same holds for X_{disc} since $0 - \max\{0.693 - 0.693, 0.686 - 0.689\} = 0$. Note however that this does not mean that CMIM always performs correctly. As discussed in Section 5.1, CMIM suffers from the problems of redundancy undervalued and complementarity penalized, and Example 3 provides a case where CMIM decides incorrectly.

ACCEPTED MANUSCRIPT

7. Conclusions

This paper carried out an evaluation and a comparison of forward feature selection methods based on mutual information. For this evaluation we selected methods representative of all types of feature selection methods proposed in the literature, namely MIM, MIFS, mRMR, maxMIFS, CIFE, JMI, CMIM, and JMIM. The evaluation was carried out theoretically, i.e. independently of the specificities of datasets and classifiers; thus, our results establish unequivocally the relative merits of the methods.

Forward feature selection methods iterate step-by-step and select one feature at each step, among the set of candidate features, the one that maximizes an objective function expressing the contribution each candidate feature to the explanation of the class. In our case, the mutual information (MI) is used as the measure of association between the class and the features. Specifically, the candidate feature selected at each step is the one that maximizes the MI between the class and the set formed by the candidate feature and the already selected features.

Our theoretical evaluation is grounded on a target objective function that the methods try to approximate and on a categorization features according to their contribution to the explanation of the class. The features are categorized as irrelevant, redundant, relevant, and fully relevant. This categorization has two novelties regarding previous works: first, we introduce the important category of fully relevant features; second, we separate non-relevant features in two categories of irrelevant and redundant features. Fully relevant features are features that fully explain the class and, therefore, its detection can be used as a stopping criterion of the feature selection process. Irrelevant and redundant features have different properties, which explains why we considered them separately. In particular, we showed that a redundant feature will always remain redundant at subsequent steps of the feature selection process, while an irrelevant feature may later turn into relevant. An important practical consequence is that redundant features, once detected, may be removed from the set

ACCEPTED MANUSCRIPT

of candidate features.

We derive upper and lower bounds for the target objective function and relate these bounds with the feature types. In particular, we showed that fully relevant features reach the maximum of the target objective function, irrelevant and redundant features reach the minimum, and relevant features take a value in between. This framework (target objective function, feature types, and objective function values for each feature type) provides a theoretical setting that can be used to compare the actual feature selection methods. Under this framework, the correct decisions at each step of the feature selection process are to select fully relevant features first and only afterwards relevant features, leave irrelevant features for future consideration (since they can later turn into relevant), and discard redundant features (since they will remain redundant).

Besides the theoretical framework, we defined a distributional setting, based on the definition of specific class, features, and a performance metric, designed to highlight the various deficiencies of methods. The setting includes four features, each belonging to one of the feature types defined above, and a class with two possible values. As performance metric, we introduced the minimum Bayes risk, a theoretical measure that does not rely on specific datasets and classifiers. The metric corresponds to the minimum total probability of misclassification for a certain class and set of selected features.

Actual feature selection methods are based on approximations of the target objective function. The target objective function comprises three terms, expressing the association between the candidate feature and the class (the relevance), the association between the candidate feature and the already selected features (the inter-feature redundancy), and the association between the candidate feature and the already selected features given the class (the class-relevant redundancy). The class-relevant redundancy is sometimes coined as the *good* redundancy, since it expresses the contribution of the candidate feature to the explanation of the class, when taken together with already selected features. We also say that this term reflects the *complementarity* between the candidate and the already selected features with respect to the class.

ACCEPTED MANUSCRIPT

Method MIM was the first method to be proposed, and completely ignored redundancy. Methods MIFS, mRMR, and maxMIFS ignored complementary effects, i.e. they did not include the class-relevant redundancy term in their objective functions. These methods lose both the upper and lower bounds of the target objective function and, more importantly, lose the connection between the bounds and the specific feature types, i.e. it is no longer possible to guarantee that fully relevant and relevant features are selected before redundant and irrelevant features, or that fully relevant come before relevant.

Methods CIFE, JMI, CMIM, and JMIM considered complementarity effects, but in different ways. The main difference between these methods lies in the approximation of the redundancy terms (the ones related with inter-feature and class-relevant redundancies). These terms depend on the complete set of already selected features and are difficult to estimate. To overcome this difficulty, the methods approximate the redundancy terms by a function of already selected features taken individually. In particular, CIFE uses the sum of the associations with the individual already selected features, JMI uses the average, and both CMIM and JMIM use the maximum. In relation to other methods, JMIM introduced an extra term in its objective function, which is unimportant and leads to confusion in the selection process. The approximations of the remaining methods lead to the following problems: CIFE overscales the redundancy, JMI undervalues the redundancy, and CMIM undervalues the redundancy in a lower extent than JMI but penalizes the complementarity. The consequences of these approximations are that CIFE loses both the upper and lower bound of the target objective function, JMI and CMIM preserve only the lower bound, and JMIM preserves only the upper bound. Moreover, as in the case of the methods that ignore complementary, the methods lose the connection between the bounds of the target objective function and the specific feature types, except in a specific case for CMIM. The drawbacks of the various methods were summarized in Table 4.

These results show that, for all methods, it is always possible to find cases where incorrect decisions are produced, and we have provided several examples

ACCEPTED MANUSCRIPT

throughout the paper and as part of our distributional setting. However, the drawbacks of the methods have different degrees of severity. MIM is a very basic method that we only considered for reference purposes. Ignoring complementary is a severe drawback that can lead to gross errors in the selection process. Thus, MIFS, mRMR, and maxMIFS, should be avoided. Regarding the methods that include complementarity effects, CIFE and JMIM should also be avoided, CIFE because its objective function is unbounded both inferiorly and superiorly due to the overscaled redundancy drawback, and JMIM because its objective function includes a bad approximation of an unimportant term that leads to confusion. Thus, the methods that currently have superior performance are JMI and CMIM. There is no clear-cut decision between these two methods, since both present drawbacks of equivalent degree of severity. JMI undervalues both the inter-feature and the class-relevant redundancy. CMIM also undervalues both types of redundancy. However, it tends to approximate better the inter-feature redundancy, but worse the class-relevant redundancy due to the problem of complementarity penalized.

Acknowledgements

This work was partially funded by Fundação para a Ciência e a Tecnologia (FCT) through project UID/Multi/04621/2013. Francisco Macedo also acknowledges the support of FCT via a PhD grant.

Appendix A. Computation of the terms in the tables of Section 6.3

In this section, we derive the expressions required for completing the tables given in Section 6.3.

Appendix A.1. Values in Table 6

Univariate differential entropies (continuous features).. The entropy of X is obtained from Example 1, in Section 2.2, considering $n = 1$. As for the entropy of the other continuous feature, $X - k'Y$, the same expression can be used since

ACCEPTED MANUSCRIPT

it is widely known that linear or affine combinations of independent univariate features following normal distributions also follow normal distributions. All we need are the variances of these two features. The variance of X is 1 and the variance of $X - k'Y$ is $1 + k'^2$. Therefore, the corresponding entropies are $h(X) = \frac{1}{2}\ln(2\pi e)$ and $h(X - k'Y) = \frac{1}{2}\ln(2\pi e(1 + k'^2))$, respectively.

Univariate entropies (discrete features and class).. Concerning Z, applying Definition 1, $H(Z) = \ln(2)$.

We now discuss the values of $H(X_{\text{disc}})$ and $H(C_k)$. Given that X and Y are independent and individually follow standard normal distributions, the joint density function of (X, Y) is

$$f_{X,Y}(x, y) = \frac{1}{2\pi}\exp(-(x^2 + y^2)) = \phi(x)\phi(y).$$

Therefore, the density at point (x, y) only depends on the distance from this point to the origin, $\sqrt{x^2 + y^2}$, in the context of the two-dimensional space defined by (X, Y). As a consequence, the probability of (X, Y) taking values in a region limited by two rays having the origin as starting point is given by $\alpha/(2\pi)$, with α denoting the angle between the two rays, as illustrated in Figure A.4. The circle is dashed in the figure since we can consider an infinite radius.

Considering an infinite radius is in fact what we need. Both C_k and X_{disc} are characterized by a partition of $\mathbb{R}^2$ in two regions separated by a line that crosses the origin. Therefore, each region covers an angle $\alpha = \pi$, so that each region has associated probability 1/2. Thus, C_k and X_{disc} follow a Bernoulli distribution with success probability 1/2, just as Z. As a result, $H(C_k) = H(X_{\text{disc}}) = H(Z) = \ln(2)$.

Appendix A.2. Values in Table 7

Continuous features.. In order to derive $\text{MI}(C_k, X)$ and $\text{MI}(C_k, X - k'Y)$, expression (10) should be, in general, preferred over (11). In fact, the entropy of a feature given the class can be obtained through the corresponding probability density functions; recall (3). These, in turn, are possible to derive easily in our

ACCEPTED MANUSCRIPT

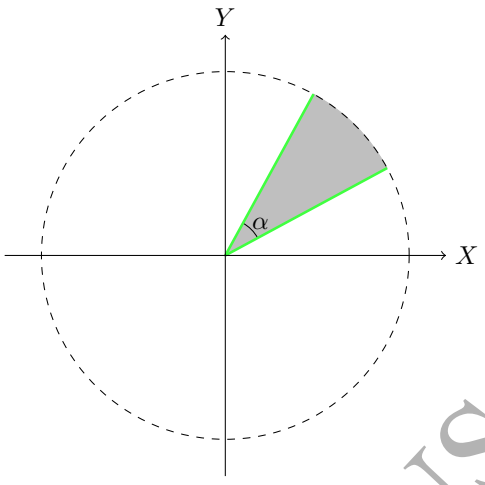


Figure A.4: Angle between two rays starting from the origin; $\alpha/(2\pi)$ is the probability that (X, Y) belongs to the region delimited by the two rays, when X and Y are two independent random variables following standard normal distribution.

setting. Therefore, we calculate the MI of interest using the representation

$$\mathrm{MI}(X_i, C_k) = h(X_i) - \sum_{j=0}^{1} \int f_{X_i|C_k=j}(u) P(C_k = j) \ln f_{X_i|C_k=j}(u) du, \quad \text{(A.1)}$$

where $P(C_k = j) = 1/2$, $j = 0, 1$. It all comes down to determining $f_{X_i|C_k=j}(u)$, $j = 0, 1$, as $h(X_i)$ is known (vide Table 6).

Given that the features follow a normal distribution, the conditional distribution of interest is the well-known skew-normal distribution [vide 23, Ch. 5]. Therefore, we only need to determine, in each case, the parameters of the mentioned distribution; recall (31).

In the case of $\mathrm{MI}(C_k, X)$, it was proved [23, Ch. 5] that $X|C_k = j$, $j = 0, 1$, follow skew-normal distributions with parameters $(0, 1, \frac{(-1)^{j+1}}{k})$; i.e. $X|C_k = j \sim \mathrm{SN}(0, 1, \frac{(-1)^{j+1}}{k})$, $j = 0, 1$.

As for $\mathrm{MI}(C_k, X - k'Y)$, we use the procedure used for the determination of $\mathrm{MI}(C_k, X)$ [23, Ch. 5] to prove that $X - k'Y|C_k = j \sim \mathrm{SN}(0, \sqrt{1 + k'^2}, (-1)^{j+1}(\frac{1-kk'}{k+k'}))$, $j = 0, 1$. The procedure consists of obtaining the conditional distribution functions of the feature given the two different possible values of the class, taking

ACCEPTED MANUSCRIPT

then the corresponding derivatives in order to obtain the associated probability density functions.

In this context, we will need the probability density function $f_{X+kY,X-k'Y}(z,w)$. This can be obtained from the joint density of the pair (X,Y). In fact, there is a way to obtain the probability density function of $g(X,Y)$, with g being a bijective function, from the probability density function of (X,Y), using the general well-known expression [49, Ch. 2]

$$f_{g(X,Y)}(z,w) = f_{X,Y}(g^{-1}(z,w)) \left| \frac{dg^{-1}(z,w)}{d(z,w)} \right|, \tag{A.2}$$

where $|\frac{dg^{-1}(z,w)}{d(z,w)}|$ denotes the absolute value of the Jacobian of the inverse of the function g.

As for the inverse function of the transformation $g(X,Y) = (X + kY, X - k'Y)$, it is given by

$$g^{-1}(z,w) = \left(\frac{kz + k'w}{k+k'}, \frac{w-z}{k+k'} \right).$$

The absolute value of its Jacobian is $|1/(k+k')|$. As both k and k' are non-negative, this can be simply written as $1/(k+k')$.

As a result, we have

$$f_{X+kY,X-k'Y}(z,w) = \frac{1}{k+k'} \phi\left(\frac{kz+k'w}{k+k'} \right) \phi\left(\frac{w-z}{k+k'} \right), \quad (z,w) \in \mathbb{R}^2.$$

We can now proceed with the derivation of the distribution functions of interest. From now on, the distribution function of $\boldsymbol{Z}$ at $\boldsymbol{z}$ will be represented by $F_{\boldsymbol{Z}}(z)$.

ACCEPTED MANUSCRIPT

We start with the case $C_k = 0$:

$$
\begin{aligned}
&F_{X-k'Y|X+kY<0}(u) = P(X - k'Y \le u | X + kY < 0) \\
&= \frac{P(X - k'Y \le u, X + kY < 0)}{P(X + kY < 0)} \\
&= 2\int_{-\infty}^{0}\int_{-\infty}^{u} f_{X+kY,X-k'Y}(z,w)dw\,dz \\
&= 2\int_{-\infty}^{0}\int_{-\infty}^{u} \frac{1}{k+k'}\phi(\frac{kz+k'w}{k+k'})\phi(\frac{w-z}{k+k'})dw\,dz \\
&= 2\int_{-\infty}^{0}\int_{-\infty}^{u} \frac{1}{k+k'}\frac{1}{2\pi}\exp\left\{-\frac{1}{2}\frac{(k^2+1)z^2}{(k+k')^2}\right\}\exp\left\{-\frac{1}{2}(k'^2+1)[w-\frac{z(1-kk')}{1+k'^2}]^2 + \right. \\
&\qquad \left. z^2[\frac{(1+k^2)(1+k'^2)-(1-kk')^2}{1+k'^2}]\right\}dw\,dz \\
&= \int_{-\infty}^{u}\frac{1}{\sqrt{\pi}}\exp\left\{-\frac{1}{2}\frac{(k+k')^2z^2}{(k+k')^2}\right\}\int_{-\infty}^{0}\frac{1}{\sqrt{\pi}(k+k')}\times \\
&\qquad \exp\{-\frac{1}{2}\frac{(1+k'^2)(w-\frac{z(1-kk')}{1+k'^2})^2}{(k+k')^2}\}dz\,dw \\
&= \sqrt{2}\int_{-\infty}^{u}\frac{1}{\sqrt{2\pi}}\frac{\sqrt{2}}{\sqrt{1+k'^2}}\int_{-\infty}^{0}\frac{1}{\frac{\sqrt{2\pi}(k+k')}{\sqrt{1+k'^2}}}\exp\left\{-\frac{1}{2}\frac{z^2(k+k')^2}{1+k'^2}\right\}dz\,dw \\
&= \frac{\sqrt{2}}{\sqrt{1+k'^2}}\int_{-\infty}^{u}\frac{1}{\sqrt{2\pi}\sqrt{1+k'^2}(k+k')}\exp\left\{-\frac{1}{2}\frac{z^2}{1+k'^2}\right\}\Phi(-\frac{(1-kk')z}{(k-k')\sqrt{1+k'^2}})dz \\
&= \int_{-\infty}^{u}\frac{2}{\sqrt{1+k'^2}}\phi(\frac{z}{\sqrt{1+k'^2}})\Phi(-\frac{(1-kk')z}{(k-k')\sqrt{1+k'^2}})dz.
\end{aligned}
$$

Hence, $X - k'Y|X + kY < 0 \sim \mathrm{SN}(0, \sqrt{1+k'^2}, \frac{-(1-kk')}{k+k'})$.

Some auxiliary steps were required in the first step of the derivation above in which $\frac{1}{k+k'}\phi(\frac{kz+k'w}{k+k'})\phi(\frac{z-w}{k-k'})$ was transformed significantly. The main technicality about such steps was the algebraic manipulation

$$
\begin{aligned}
&(kz+k'w)^2 + (w-z)^2 \\
&= (1+k^2)z^2 - 2wz(1-kk') + (k'^2+1)w^2 \\
&= (1+k'^2)\{w^2 - 2w\frac{z(1-kk')}{1+k'^2} + [\frac{z(1-kk')}{1+k'^2}]^2 - [\frac{z(1-kk')}{1+k'^2}]^2\} + (1+k^2)z^2 \\
&= (1+k'^2)\{[w-\frac{z(1-kk')}{1+k'^2}]^2 - \frac{[z(1-kk')]^2}{(1+k'^2)^2}\} + (1+k^2)z^2 \\
&= (1+k'^2)[w-\frac{z(1-kk')}{1+k'^2}]^2 + z^2\frac{(k+k')^2}{1+k'^2}.
\end{aligned}
$$

ACCEPTED MANUSCRIPT

As for the conditional case in which $C_k = 1$, we provide a briefer version of the computation as most steps are the same as for $C_k = 0$.

$$
\begin{aligned}
&F_{X-k'Y|X+kY\geq 0}(u) \\
&= P(X - k'Y \leq u | X + kY \geq 0) \\
&= \frac{P(X - k'Y \leq u, X + kY \geq 0)}{P(X + kY \geq 0)} \\
&= 2\int_0^{+\infty}\int_{-\infty}^{u} f_{X+kY,X-k'Y}(z,w)dw\,dz \\
&= 2\int_0^{+\infty}\int_{-\infty}^{u} \frac{1}{k+k'}\phi(\frac{kz+k'w}{k+k'})\phi(\frac{w-z}{k+k'})dw\,dz \\
&= 2\int_{-\infty}^{0}\int_{-\infty}^{u} \frac{1}{k+k'}\frac{1}{2\pi}\exp\left\{-\frac{1}{2}\frac{(k^2+1)z^2}{(k+k')^2}\right\}\exp\left\{-\frac{1}{2}(k'^2+1)[w-\frac{z(1-kk')}{1+k'^2}]^2 + \right. \\
&\qquad \left. z^2[\frac{(1+k^2)(1+k'^2)-(1-kk')^2}{1+k'^2}]\right\}dw\,dz \\
&= \int_{-\infty}^{u} \frac{2}{\sqrt{1+k'^2}}\phi(\frac{z}{\sqrt{1+k'^2}})[1-\Phi(-\frac{(1-kk')z}{(k-k')\sqrt{1+k'^2}})]dz.
\end{aligned}
$$

Given the symmetry of the normal distribution, we know that $(1 - \Phi(-x)) = \Phi(x)$. This allows reducing the expression to

$$\int_{-\infty}^{u} \frac{2}{\sqrt{1+k'^2}}\phi(\frac{z}{\sqrt{1+k'^2}})\Phi(\frac{(1+kk')z}{(k-k')\sqrt{1+k'^2}})dz.$$

Thus, $X - k'Y|X + kY \geq 0 \sim \text{SN}\left(0, \sqrt{1+k'^2}, \frac{(1-kk')}{k+k'}\right)$.

Discrete features.. In order to obtain $\text{MI}(C_k, X_{\text{disc}})$, we use (7). We have $\text{MI}(C_k, X_{\text{disc}}) = H(C_k) + H(X_{\text{disc}}) - H(C_k, X_{\text{disc}})$. We only need to compute $H(C_k, X_{\text{disc}})$ since the required univariate entropies are known from Table 6. From Definition 1, this requires obtaining the probabilities of the four possible combinations of values associated with the pair (C_k, X_{disc}). We represent the regions associated with such values in Figure A.5, considering the two-dimensional space defined by the pair (X, Y). Considering the reasoning used to obtain $H(C_k)$ and $H(X_{\text{disc}})$, associated with Figure A.4, we only need the four angles covered by the associated four regions in order to compute their corresponding probabilities. The determination of such angles only requires the

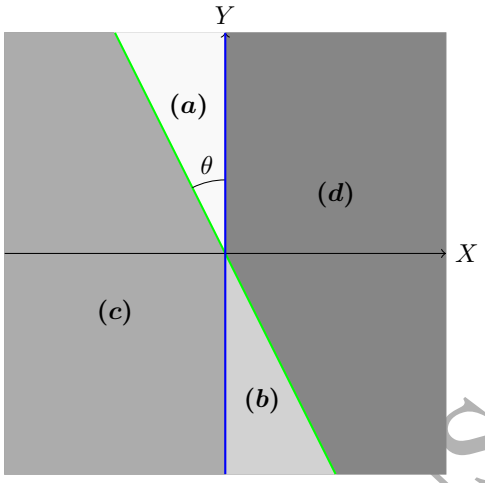


Figure A.5: Four regions associated with the joint distribution of (X_{disc}, C_k), where $\theta = \arctan k$, when X and Y are two independent random variables following standard normal distribution.

knowledge of θ, represented in Figure A.5 since the remaining angles consist of its supplementary, its opposite, and the opposite of its supplementary.

In the end, there are two angles (associated with regions (a) and (b) in Figure A.5) whose value is $\arctan k$, while the other two (associated with regions (c) and (d) in Figure A.5) have the value $(\pi - \arctan k)$, implying that

$$P(X_{\text{disc}} = u, C_k = j) = \begin{cases} \frac{\pi - \arctan k}{2\pi}, & u = 0, j = 0 \text{ and } u = 1, j = 1 \\ \frac{\arctan k}{2\pi}, & u = 0, j = 1 \text{ and } u = 1, j = 0 \end{cases}. \quad \text{(A.3)}$$

As a result,

$$H(C_k, X_{\text{disc}}) = -2\times\left(\frac{\arctan k}{2\pi}\ln(\frac{\arctan k}{2\pi}) + (\frac{1}{2} - \frac{\arctan k}{2\pi})\ln(\frac{1}{2} - \frac{\arctan k}{2\pi})\right).$$

We obtain

$$\text{MI}(C_k, X_{\text{disc}}) = 2\ln(2) + \frac{\arctan k}{\pi}\ln(\frac{\arctan k}{2\pi}) + (1 - \frac{\arctan k}{\pi})\ln(\frac{1}{2} - \frac{\arctan k}{2\pi}).$$

As for $\text{MI}(C_k, Z)$, its value is 0 since C_k and Z are independent.

ACCEPTED MANUSCRIPT

Appendix A.3. Values in Table 8

We start with $\mathrm{MI}(X, X_{\mathrm{disc}})$. X_{disc} is redundant given $\{X\}$, so that $H(X_{\mathrm{disc}}|X) = 0$. Additionally, $H(X_{\mathrm{disc}}) = \ln(2)$ (vide Table 6), so that, by (4), $\mathrm{MI}(X, X_{\mathrm{disc}}) = H(X_{\mathrm{disc}}) - H(X_{\mathrm{disc}}|X) = \ln(2) - 0 = \ln(2)$.

As for $\mathrm{MI}(X - k'Y, X_{\mathrm{disc}})$, we first note that the deduction of $\mathrm{MI}(C_k, X)$ [23, Ch. 5] can be similarly done for the case where k in (28) is allowed to be negative, from which one would conclude that $\mathrm{MI}(C_{-k'}, X) = \mathrm{MI}(C_{k'}, X)$. In turn, X_{disc} and $X - k'Y$ are obtained from $C_{k'}$ and X, respectively, by rotating k' degrees anticlockwise, considering the two-dimensional space defined by the pair (X, Y). As a result, $\mathrm{MI}(X_{\mathrm{disc}}, X - k'Y) = \mathrm{MI}(C_{k'}, X)$. In fact, MI is invariant under one-to-one transformations [50, for more details]. Finally, by transitivity, $\mathrm{MI}(X - k'Y, X_{\mathrm{disc}}) = \mathrm{MI}(C_{k'}, X)$.

As for $\mathrm{MI}(X, X - k'Y)$, we use (7). We have $\mathrm{MI}(X, X - k'Y) = h(X) + h(X - k'Y) - h(X, X - k'Y)$. We only need to compute $h(X, X - k'Y)$ since the univariate entropies are known (vide Table 6). As both features follow normal distributions, the joint distribution is a bivariate normal distribution, whose entropy depends on the determinant of the covariance matrix, as described in Example 1. The value of the mentioned determinant is k'^2, so that $\mathrm{MI}(X, X - k'Y) = \frac{1}{2}\ln\left(1 + 1/k'^2\right)$.

The MI terms involving Z do not require any calculation given that Z is independent of X, X_{disc}, and $X - k'Y$, implying that $\mathrm{MI}(Z, X) = \mathrm{MI}(Z, X_{\mathrm{disc}}) = \mathrm{MI}(Z, X - k'Y) = 0$.

Appendix A.4. Values in Table 9

For deriving $\mathrm{MI}(X, X_{\mathrm{disc}}|C_k)$, we use (12). We have $\mathrm{MI}(X, X_{\mathrm{disc}}|C_k) = H(X_{\mathrm{disc}}|C_k) - H(X_{\mathrm{disc}}|X, C_k) = 0$. Noting that $H(X_{\mathrm{disc}}|X, C_k) = 0$ since $H(X_{\mathrm{disc}}|X) = 0$, using (16), we have $\mathrm{MI}(X, X_{\mathrm{disc}}|C_k) = H(X_{\mathrm{disc}}|C_k)$. In turn, $H(X_{\mathrm{disc}}|C_k) = H(X_{\mathrm{disc}}) - \mathrm{MI}(X_{\mathrm{disc}}, C_k)$, where the values $H(X_{\mathrm{disc}})$ and $\mathrm{MI}(X_{\mathrm{disc}}, C_k)$ have been obtained; recall Tables 6 and 8, respectively. We conclude that

$$\mathrm{MI}(X, X_{\mathrm{disc}}|C_k) = -\frac{\arctan k}{\pi}\ln\left(\frac{\arctan k}{\pi}\right) - \left(1 - \frac{\arctan k}{\pi}\right)\ln\left(1 - \frac{\arctan k}{\pi}\right).$$

ACCEPTED MANUSCRIPT

Concerning $\mathrm{MI}(X, X-k'Y|C_k)$, we use (13). We have $\mathrm{MI}(X, X-k'Y|C_k) = h(X|C_k)+h(X-k'Y|C_k)-h(X, X-k'Y|C_k)$. The first two terms were obtained in the context of the determination of $\mathrm{MI}(X, C_k)$ and $\mathrm{MI}(X - k'Y|C_k)$; they consist of the second term in (A.1).

As in the computation of class-conditional entropies for univariate continuous features, recall (A.1), we only need the joint probability density functions of $(X, X - k'Y)$ conditioned on the two possible values of the class to determine $h(X, X-k'Y|C_k)$. These are obtained from the corresponding conditional probability density functions associated with the pair (X, Y), using (A.2).

We first need to derive the probability density functions associated with the distributions $(X, Y)|C_k = j$, $j = 0, 1$. These are obtained starting from the corresponding distribution functions, as done in the context of the determination of the probability density functions needed in sequence of (A.1).

We start with $C_k = 1$. We have

$$\begin{aligned} F_{(X,Y)|C_k=1}(x, y) =& \frac{P(X \leq x, Y \leq y, X + kY \geq 0)}{P(X + kY \geq 0)} \\ =& 2P(X \leq x, Y \leq y, X + kY \geq 0). \end{aligned}$$

In order to proceed, we separate the expression in two cases. In fact, if $x < -ky$, the value is simply 0. If, instead, $x \geq -ky$, we have

$$\begin{aligned} 2P(X \leq x, Y \leq y, X + kY \geq 0) =& 2 \int_{-\infty}^{x} \int_{-\frac{u}{k}}^{y} \phi(u)\phi(v) dv\, du \\ =& 2 \int_{-\infty}^{x} \phi(u)[\Phi(v) - \Phi(-\frac{u}{k})] du \\ =& 2\Phi(y)\Phi(x) - F_{\mathrm{SN}(0,1,-\frac{1}{k})}(x). \end{aligned}$$

Thus, the corresponding density is given by

$$f_{(X,Y)|C_k=1}(x, y) = \begin{cases} 2\phi(y)\phi(x), & x \geq -ky \\ 0, & x < -ky \end{cases}.$$

We now make the same type of deduction for $C_k = 0$. We have

$$\begin{aligned} F_{(X,Y)|C_k=0}(x, y) =& \frac{P(X \leq x, Y \leq y, X + kY < 0)}{P(X + kY < 0)} \\ =& 2P(X \leq x, Y \leq y, X + kY < 0). \end{aligned}$$

ACCEPTED MANUSCRIPT

If $x < -ky$, we obtain

$$\begin{aligned}2P(X \leq x, Y \leq y, X + kY < 0) =& 2\int_{-\infty}^{x}\int_{-\infty}^{y} \phi(u)\phi(v)dv\,du \\ =& 2\Phi(y)\Phi(x).\end{aligned}$$

If, instead, $x \geq -ky$, we have

$$\begin{aligned}2P(X \leq x, Y \leq y, X + kY \geq 0) =& 2\left[\int_{-\infty}^{x}\int_{-\infty}^{y} \phi(u)\phi(v)du\,dv - \int_{-\infty}^{x}\int_{-\frac{u}{k}}^{y} \phi(u)\phi(v)dv\,du\right] \\ =& 2\Phi(y)\Phi(x) - [2\Phi(y)\Phi(x) - F_{\mathrm{SN}(0,1,-\frac{1}{k})}(x)] \\ =& F_{\mathrm{SN}(0,1,-\frac{1}{k})}(x).\end{aligned}$$

Thus, the corresponding density is given by

$$f_{(X,Y)|C_k=0}(x,y) = \begin{cases} 0, & x \geq -ky \\ 2\phi(y)\phi(x), & x < -ky \end{cases}.$$

We can now obtain $f_{X,X-k'Y}(u,v)$ using (A.2). In this case, $g^{-1}(z,w) = (z, \frac{z-w}{k'})$. The determinant of the corresponding Jacobian is $1/k'$. Therefore, we have $f_{(X,X-k'Y)|C_k=c}(u,v) = f_{(X,Y)|C_k=c}(u, \frac{u-v}{k'})\frac{1}{k'}$, $c = 0, 1$. Thus, we obtain

$$f_{(X,X-k'Y)|C_k=1}(u,v) = \begin{cases} 2\phi(u)\phi(\frac{v-u}{k'})\frac{1}{k'}, & u > \frac{k}{k'+k}v \\ 0, & u \leq \frac{k}{k'+k}v \end{cases}$$

and

$$f_{(X,X-k'Y)|C_k=0}(u,v) = \begin{cases} 0, & u > \frac{k}{k'+k}v \\ 2\phi(u)\phi(\frac{v-u}{k'})\frac{1}{k'}, & u \leq \frac{k}{k'+k}v \end{cases},$$

where we note that $\phi(\frac{v-u}{k'})\frac{1}{k'}$ can be also seen as the density for a normal distribution with parameters (u, k'^2) evaluated at v.

We have

$$h(X, X-k'Y|C_k = 1) = -\int_{-\infty}^{+\infty}\int_{\frac{k}{k'+k}v}^{+\infty} 2\phi(u)\phi(\frac{v-u}{k'})\frac{1}{k'}\ln(2\phi(u)\phi(\frac{v-u}{k'})\frac{1}{k'})dv\,du$$

and

$$h(X, X-k'Y|C_k = 0) = -\int_{-\infty}^{+\infty}\int_{-\infty}^{\frac{k}{k'+k}v} 2\phi(u)\phi(\frac{v-u}{k'})\frac{1}{k'}\ln(2\phi(u)\phi(\frac{v-u}{k'})\frac{1}{k'})dv\,du.$$

ACCEPTED MANUSCRIPT

This implies that

$$h(X, X - k'Y|C_k) = -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \phi(u)\phi(\frac{v-u}{k'})\frac{1}{k'}\ln(2\phi(u)\phi(\frac{v-u}{k'})\frac{1}{k'})dv\,du.$$

The part inside the logarithm can be re-written considering the explicit expression of the density of a standard normal distribution. We have

$$\begin{aligned}\ln(2\phi(u)\phi(\frac{v-u}{k'})\frac{1}{k'}) &= \ln(\frac{1}{\pi k'}\exp(-\frac{1}{2k'^2}[(1+k')u^2 - 2uv + v^2])) \\ &= -\ln(\pi k') - \frac{1}{2k'^2}[(1+k')u^2 - 2uv + v^2].\end{aligned}$$

We can still write this as

$$h(X, X-k'Y|C_k) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \phi(u)\phi(v)[\ln(\pi k') + \frac{1}{2k'^2}[(1+k')u^2 - 2uv + v^2]]dv\,du.$$

The final result is $\ln(\pi k') + \frac{1}{2} + \frac{1}{2} = \ln\pi + \ln k' + 1$. In fact, the first term is a double integral in the whole space of a probability density function, associated with (X, Y), times a constant, so that the result is such constant, $\ln(\pi k')$; while the remaining terms are also easy to obtain since they consist of constants multiplied with first and second order moments from the normal distribution with parameters (u, k').

As for $\mathrm{MI}(X - k'Y, X_{\mathrm{disc}}|C_k)$, we compute it, using (12), through its representation $\mathrm{MI}(X - k'Y, X_{\mathrm{disc}}|C_k) = h(X - k'Y|C_k) - h(X - k'Y|X_{\mathrm{disc}}, C_k)$. Note that $h(X - k'Y|C_k)$ has already been derived, in the context of the determination of $\mathrm{MI}(X - k'Y, C_k)$; recall (A.1). Thus, we only need to derive $h(X - k'Y|X_{\mathrm{disc}}, C_k)$.

We need to obtain $f_{X-k'Y|X_{\mathrm{disc}}=u,C_k=j}(v)$ for each possible combination of pairs (u, j). We first note that

$$\begin{aligned}f_{X-k'Y|X_{\mathrm{disc}}=u,C_k=j}(v) &= \frac{d}{dv}F_{X-k'Y|X_{\mathrm{disc}}=u,C_k=j}(v) \\ &= \frac{d}{dv}\frac{P(X - k'Y \le v, X_{\mathrm{disc}} = u, C_k = j)}{P(X_{\mathrm{disc}} = u, C_k = j)} \\ &= \frac{1}{P(X_{\mathrm{disc}} = u, C_k = j)}\frac{d}{dv}P(X - k'Y \le v, X_{\mathrm{disc}} = u, C_k = j).\end{aligned}$$

The values $P(X_{\mathrm{disc}} = u, C_k = j)$, $u = 0, 1$ and $j = 0, 1$, can be found in (A.3).

ACCEPTED MANUSCRIPT

Therefore, we only need to obtain $P(X - k'Y \leq v, X_{\text{disc}} = u, C_k = j)$. We start with $u = 1$ and $j = 1$. We have to split in two cases. For $v \geq 0$,

$$\begin{aligned}
&P(X - k'Y \leq v, X_{\text{disc}} = 1, C_k = 1) \\
&= \int_{-\frac{v}{k'+k}}^{0} \int_{-kz}^{v+k'z} \phi(w)\phi(z)dw\,dz + \int_{0}^{+\infty} \int_{0}^{v+k'z} \phi(w)\phi(z)dz\,dw \\
&= \int_{-\frac{z}{k'+k}}^{0} \phi(z)[\Phi(v + k'z) - \Phi(-kz)]dz + \int_{0}^{+\infty} \phi(z)[\Phi(v + k'z) - \frac{1}{2}]dz \\
&= \int_{-\frac{v}{k'+k}}^{0} \phi(z)\Phi(v + k'z)dz - \frac{1}{2}[F_{\text{SN}(0,1,-k)}(0) - F_{\text{SN}(0,1,-k)}(-\frac{v}{k' + k})] \\
&\qquad + \int_{0}^{+\infty} \phi(z)\Phi(v + k'z)dz - \frac{1}{4} \\
&= \int_{-\frac{v}{k'+k}}^{+\infty} \phi(z)\Phi(v + k'z)dz - \frac{1}{2}F_{\text{SN}(0,1,-k)}(0) + \frac{1}{2}F_{\text{SN}(0,1,-k)}(-\frac{v}{k' + k}) - \frac{1}{4}.
\end{aligned}$$

In turn, for $v < 0$,

$$\begin{aligned}
&P(X - k'Y \leq v, X_{\text{disc}} = 1, C_k = 1) \\
&= \int_{-\frac{v}{k'}}^{+\infty} \int_{0}^{v+k'z} \phi(w)\phi(z)dz\,dw \\
&= \int_{-\frac{v}{k'}}^{+\infty} \phi(z)[\Phi(v + k'z) - \frac{1}{2}]dz \\
&= \int_{-\frac{v}{k'}}^{+\infty} \phi(z)\Phi(v + k'z)dz - \Phi(\frac{v}{k'}).
\end{aligned}$$

We now need to take the derivative of the two expressions with respect to v to obtain the corresponding conditional density functions. In the case of $v \geq 0$,

$$\begin{aligned}
&\frac{d}{dv}\left[\int_{-\frac{v}{k'+k}}^{+\infty} \phi(z)\Phi(v + k'z)dz - \frac{1}{2}F_{\text{SN}(0,1,-k)}(0) + \frac{1}{2}F_{\text{SN}(0,1,-k)}(-\frac{v}{k' + k}) - \frac{1}{4}\right] \\
&= \int_{-\frac{v}{k'+k}}^{+\infty} \phi(z)\phi(v + k'z)dz + \frac{1}{2}f_{\text{SN}(0,1,-k)}(-\frac{v}{k' + k})\frac{1}{k' + k} \\
&\qquad - \frac{1}{2}f_{\text{SN}(0,1,-k)}(-\frac{v}{k' + k})\frac{1}{k' + k} \\
&= \int_{-\frac{v}{k'+k}}^{+\infty} \phi(z)\phi(v + k'z)dz;
\end{aligned}$$

ACCEPTED MANUSCRIPT

while, for $v < 0$,

$$\begin{aligned}
&\frac{d}{dv}\left[\int_{-\frac{v}{k'}}^{+\infty}\phi(z)\Phi(v+k'z)dz - \frac{1}{2}\Phi(\frac{v}{k'})\right] \\
&= \int_{-\frac{v}{k'}}^{+\infty}\phi(z)\phi(v+k'z)dz + \frac{1}{2}\phi(\frac{v}{k'})\frac{1}{k'} - \frac{1}{2}\phi(\frac{v}{k'})\frac{1}{k'} \\
&= \int_{-\frac{v}{k'}}^{+\infty}\phi(z)\phi(v+k'z)dz.
\end{aligned}$$

Note that the following important result [51, Ch. 3] was required in order to obtain both final expressions above:

$$\frac{d}{dx}\int_{a(x)}^{b(x)} g(x,y)dy = \int_{a(x)}^{b(x)} \frac{dg(x,y)}{dx}dy + g(x,b(x))b'(x) - g(x,a(x))a'(x). \quad \text{(A.4)}$$

This result was applied to $\frac{d}{dv}\int_{-\frac{v}{k'+k}}^{+\infty}\phi(z)\Phi(v+k'z)dz$, in the expression for $v \geq 0$, and also to $\frac{d}{dv}\int_{-\frac{v}{k'}}^{+\infty}\phi(z)\Phi(v+k'z)dz$, concerning the case $v < 0$.

The desired probability density function is

$$f_{X-k'Y|X_{\text{disc}}=1,C_k=1}(v) = \begin{cases} \frac{2\pi}{\pi-\arctan k}\int_{-\frac{v}{k'+k}}^{+\infty}\phi(z)\phi(v+k'z)dz, & v \geq 0 \\ \frac{2\pi}{\pi-\arctan k}\int_{-\frac{v}{k'}}^{+\infty}\phi(z)\phi(v+k'z)dz, & v < 0 \end{cases}.$$

We now consider $u = 0$ and $j = 1$. We have to split again in two cases. For $v \geq 0$,

$$\begin{aligned}
P(X - k'Y \leq v, X_{\text{disc}} = 0, C_k = 1) &= \int_0^{+\infty}\int_{-kz}^{0}\phi(w)\phi(z)dz\,dw \\
&= \int_0^{+\infty}\phi(z)[\frac{1}{2} - \Phi(-kz)]dz \\
&= \frac{1}{2}F_{\text{SN}(0,1,-k)}(0) - \frac{1}{4}.
\end{aligned}$$

ACCEPTED MANUSCRIPT

For $v < 0$,

$$
\begin{aligned}
&P(X - k'Y \le v, X_{\text{disc}} = 0, C_k = 1) \\
&= \int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}} \int_{-kz}^{v+k'z} \phi(w)\phi(z)dw\,dz + \int_{-\frac{v}{k'}}^{+\infty} \int_{-kz}^{0} \phi(w)\phi(z)dz\,dw \\
&= \int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}} \phi(z)[\Phi(v+k'z) - \Phi(-kz)]dz + \int_{-\frac{v}{k'}}^{+\infty} \phi(z)[\frac{1}{2} - \Phi(-kz)]dz\,dw \\
&= \int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}} \phi(z)\Phi(v+k'z)dz - \frac{1}{2}[F_{\text{SN}(0,1,-k)}(-\frac{v}{k'}) - F_{\text{SN}(0,1,-k)}(-\frac{v}{k'+k})] + \frac{1}{2}[1 - \Phi(-\frac{v}{k'})] \\
&\qquad - \frac{1}{2}[1 - F_{\text{SN}(0,1,-k)}(-\frac{v}{k'})] \\
&= \int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}} \phi(z)\Phi(v+k'z)dz + \frac{1}{2}F_{\text{SN}(0,1,-k)}(-\frac{v}{k'+k}) + \frac{1}{2}\Phi(\frac{v}{k'}).
\end{aligned}
$$

We now need to take the derivative of the two expressions with respect to v to obtain the corresponding conditional density functions. In the case of $v \ge 0$, it is simply 0 since there is no dependency on v. As for $v < 0$,

$$
\begin{aligned}
&\frac{d}{dv}\left[\int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}} \phi(z)\Phi(v+k'z)dz + \frac{1}{2}F_{\text{SN}(0,1,-k)}(-\frac{v}{k'+k}) + \frac{1}{2}\Phi(\frac{v}{k'})\right] \\
&= \int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}} \phi(z)\phi(v+k'z)dz + \frac{1}{2}f_{\text{SN}(0,1,-k)}(-\frac{v}{k'+k})\frac{1}{k'+k} - \frac{1}{2}\phi(\frac{v}{k'})\frac{1}{k'} \\
&\qquad - \frac{1}{2}f_{\text{SN}(0,1,-k)}(-\frac{v}{k'+k})\frac{1}{k'+k} + \frac{1}{2}\phi(\frac{v}{k'})\frac{1}{k'} \\
&= \int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}} \phi(z)\phi(v+k'z)dz.
\end{aligned}
$$

Once again, (A.4) was applied, in this case to $\int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}} \phi(z)\Phi(v+k'z)dz$.

The desired probability density function is

$$
f_{X-k'Y|X_{\text{disc}}=0,C_k=1}(v) = \begin{cases} 0, & v \ge 0 \\ \frac{2\pi}{\arctan k}\int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}} \phi(z)\phi(v+k'z)dz, & v < 0 \end{cases}.
$$

ACCEPTED MANUSCRIPT

We now consider $u = 1$ and $j = 0$. For $v \geq 0$,

$$
\begin{aligned}
&P(X - k'Y \leq v, X_{\text{disc}} = 1, C_k = 0) \\
&= \int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}} \int_0^{v+k'z} \phi(w)\phi(z)dw\,dz + \int_{-\frac{v}{k'+k}}^{0} \int_0^{-kz} \phi(w)\phi(z)dz\,dw \\
&= \int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}} \phi(z)[\Phi(v+k'z) - \frac{1}{2}]dz + \int_{-\frac{v}{k'+k}}^{0} \phi(z)[\Phi(-kz) - \frac{1}{2}]dz\,dw \\
&= \int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}} \phi(z)\Phi(v+k'z)dz - \frac{1}{2}[\Phi(-\frac{v}{k'+k}) - \Phi(-\frac{v}{k'})] \\
&\qquad + \frac{1}{2}[F_{\text{SN}(0,1,-k)}(0) - F_{\text{SN}(0,1,-k)}(-\frac{v}{k'+k})] - \frac{1}{2}\Phi(\frac{v}{k'+k}) \\
&= \int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}} \phi(z)\Phi(v+k'z)dz - \frac{1}{2}\Phi(\frac{v}{k'+k}) \\
&\qquad + \frac{1}{2}F_{\text{SN}(0,1,-k)}(0) - \frac{1}{2}F_{\text{SN}(0,1,-k)}(-\frac{v}{k'+k}).
\end{aligned}
$$

As for the case $v < 0$, $P(X - k'Y \leq v, X_{\text{disc}} = 1, C_k = 0) = 0$.

We now need to take the derivative with respect to v of the expression obtained for $v \geq 0$ (the derivative of the one for $v < 0$ is 0) to obtain the corresponding conditional density functions. We have

$$
\begin{aligned}
&\frac{d}{dv}\left[\int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}} \phi(z)\Phi(v+k'z)dz - \frac{1}{2}\Phi(-\frac{v}{k'+k}) + \frac{1}{2}F_{\text{SN}(0,1,-k)}(0) - \frac{1}{2}F_{\text{SN}(0,1,-k)}(-\frac{v}{k'+k})\right] \\
&= \int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}} \phi(z)\phi(v+k'z)dz - \frac{1}{2}\phi(-\frac{v}{k'+k}) - \frac{1}{2}f_{\text{SN}(0,1,-k)}(-\frac{v}{k'+k})\frac{1}{k'+k} \\
&\qquad + \frac{1}{2}\phi(-\frac{v}{k'+k}) + \frac{1}{2}f_{\text{SN}(0,1,-k)}(-\frac{v}{k'+k})\frac{1}{k'+k} \\
&= \int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}} \phi(z)\phi(v+k'z)dz.
\end{aligned}
$$

We again applied (A.4), in this case to $\int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}} \phi(z)\Phi(v+k'z)dz$.

The desired probability density function is

$$
f_{X-k'Y|X_{\text{disc}}=1,C_k=0}(v) = \begin{cases} \frac{2\pi}{\pi - \arctan k} \int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}} \phi(z)\phi(v+k'z)dz, & v \geq 0 \\ 0, & v < 0 \end{cases}.
$$

ACCEPTED MANUSCRIPT

We finally consider $u = 0$ and $j = 0$. For $v \geq 0$,

$$\begin{aligned}
&P(X - k'Y \leq v, X_{\text{disc}} = 0, C_k = 0) \\
&= \int_{-\infty}^{0} \int_{-\infty}^{+\infty} \phi(w)\phi(z)dw\, dz - \int_{0}^{+\infty} \int_{-kz}^{0} \phi(w)\phi(z)dz\, dw \\
&\qquad - \int_{-\infty}^{-\frac{v}{k'}} \int_{v+k'z}^{0} \phi(w)\phi(z)dz\, dw \\
&= \frac{1}{2} - \int_{-\infty}^{0} \phi(z) \left[\Phi(-kz) - \frac{1}{2}\right] dz - \int_{-\infty}^{-\frac{v}{k'}} \left[\frac{1}{2} - \Phi(v + k'z)\right] \phi(z)dz \\
&= \frac{3}{4} - \frac{1}{2} F_{\text{SN}(0,1,-k)}(0) - \frac{1}{2}\Phi\left(-\frac{v}{k'}\right) + \int_{-\infty}^{-\frac{v}{k'}} \Phi(v + k'z)\phi(z)dz.
\end{aligned}$$

As for the case $v < 0$,

$$\begin{aligned}
&P(X - k'Y \leq v, X_{\text{disc}} = 0, C_k = 0) \\
&= \int_{-\infty}^{-\frac{v}{k'+k}} \int_{-\infty}^{v+k'z} \phi(w)\phi(z)dw\, dz + \int_{-\frac{v}{k'+k}}^{+\infty} \int_{-\infty}^{-kz} \phi(w)\phi(z)dz\, dw \\
&= \int_{-\infty}^{-\frac{v}{k'+k}} \Phi(v + k'z)\phi(z)dw\, dz + \int_{-\frac{v}{k'+k}}^{+\infty} \Phi(-kz)\phi(z)dz \\
&= \int_{-\infty}^{-\frac{v}{k'+k}} \Phi(v + k'z)\phi(z)dz + \frac{1}{2} - \frac{1}{2} F_{\text{SN}(0,1,-k)}\left(-\frac{v}{k'+k}\right).
\end{aligned}$$

We again need to take the derivative of the two expressions with respect to v to obtain the corresponding conditional density functions. In the case of $v \geq 0$,

$$\begin{aligned}
&\frac{d}{dv}\left[\frac{1}{4} + \frac{1}{2} F_{\text{SN}(0,1,k)}(0) - \frac{1}{2}\Phi(-\frac{v}{k'}) + \int_{-\infty}^{\frac{v}{k'}} \Phi(v + k'z)\phi(z)dz\right] \\
&= \int_{-\infty}^{\frac{v}{k'}} \phi(v + k'z)\phi(z)dz - \frac{1}{2}\phi(-\frac{v}{k'})\frac{1}{k'} + \frac{1}{2}\phi(-\frac{v}{k'})\frac{1}{k'} \\
&= \int_{-\infty}^{\frac{v}{k'}} \phi(z)\phi(v + k'z)dz;
\end{aligned}$$

ACCEPTED MANUSCRIPT

while, for $v < 0$,

$$\begin{aligned}
&\frac{d}{dv}\left[\int_{-\infty}^{-\frac{v}{k'+k}} \Phi(v+k'z)\phi(z)dz + \frac{1}{2} - \frac{1}{2}F_{\mathrm{SN}(0,1,-k)}(-\frac{v}{k'+k})\right] \\
&= \int_{-\infty}^{-\frac{v}{k'+k}} \phi(z)\phi(v+k'z)dz - \frac{1}{2}f_{\mathrm{SN}(0,1,-k)}(-\frac{v}{k'+k})\frac{v}{k'+k} \\
&\qquad + \frac{1}{2}f_{\mathrm{SN}(0,1,-k)}(-\frac{v}{k'+k})\frac{v}{k'+k} \\
&= \int_{-\infty}^{-\frac{v}{k'+k}} \phi(z)\phi(v+k'z)dz.
\end{aligned}$$

We applied (A.4) to $\int_{-\infty}^{\frac{v}{k'}} \Phi(v+k'z)\phi(z)dz$ and $\int_{-\infty}^{-\frac{v}{k'+k}} \Phi(v+k'z)\phi(z)dz$.

The desired probability density function is

$$f_{X-k'Y|X_{\mathrm{disc}}=0,C_k=0}(v) = \begin{cases} \frac{2\pi}{\pi-\arctan k}\int_{-\infty}^{\frac{v}{k'}} \phi(z)\phi(v+k'z)dz, & v \geq 0 \\ \frac{2\pi}{\pi-\arctan k}\int_{-\infty}^{-\frac{v}{k'+k}} \phi(z)\phi(v+k'z)dz, & v < 0 \end{cases}.$$

We can finally obtain an expression for $h(X-k'Y|X_{\mathrm{disc}}, C_k)$:

$$\begin{aligned}
&-\frac{\pi-\arctan k}{2\pi}\Bigg(\int_{-\infty}^{0}(\int_{-\frac{v}{k'}}^{+\infty}\frac{2\pi}{\pi-\arctan k}\zeta(z,v)dz)\ln(\int_{-\frac{v}{k'}}^{+\infty}\frac{2\pi}{\pi-\arctan k}\zeta(z,v)dz)dv + \\
&\quad\int_{0}^{+\infty}(\int_{-\frac{v}{k'+k}}^{+\infty}\frac{2\pi}{\pi-\arctan k}\zeta(z,v)dz)\ln(\int_{-\frac{v}{k'+k}}^{+\infty}\frac{2\pi}{\pi-\arctan k}\zeta(z,v)dz)dv + \\
&\quad\int_{-\infty}^{0}(\int_{-\infty}^{-\frac{v}{k'+k}}\frac{2\pi}{\pi-\arctan k}\zeta(z,v)dz)\ln(\int_{-\infty}^{-\frac{v}{k'+k}}\frac{2\pi}{\pi-\arctan k}\zeta(z,v)dz)dv + \\
&\quad\int_{0}^{+\infty}(\int_{-\infty}^{\frac{v}{k'}}\frac{2\pi}{\pi-\arctan k}\zeta(z,v)dz)\ln(\int_{-\infty}^{\frac{v}{k'}}\frac{2\pi}{\pi-\arctan k}\zeta(z,v)dz)dv\Bigg) \\
&-\frac{\arctan k}{2\pi}\Bigg(\int_{-\infty}^{0}(\int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}}\frac{2\pi}{\arctan k}\zeta(z,v)dz)\ln(\int_{-\frac{v}{k'+k}}^{-\frac{v}{k'}}\frac{2\pi}{\arctan k}\zeta(z,v)dz)dv + \\
&\quad\int_{0}^{+\infty}(\int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}}\frac{2\pi}{\arctan k}\zeta(z,v)dz)\ln(\int_{-\frac{v}{k'}}^{-\frac{v}{k'+k}}\frac{2\pi}{\arctan k}\zeta(z,v)dz)dv\Bigg),
\end{aligned} \tag{A.5}$$

where $\zeta(z,v)$ is the function $\phi(v+k'z)\phi(z)$.

As for the class-conditional MI values that involve Z, $\mathrm{MI}(Z, X|C_k) = \mathrm{MI}(Z, X-k'Y|C_k) = \mathrm{MI}(Z, X_{\mathrm{disc}}|C_k) = 0$. This requires checking that pairwise class-conditional independence holds for the three involved pairs. This follows, as

ACCEPTED MANUSCRIPT

argued in Section 6.2, from the fact that Z is independent of the pair composed by C_k and any other input feature.

Appendix B. Calculations of MBR values in Section 6.4

In this section, we start by obtaining the value of $\mathrm{MBR}(C_k, \{X\})$. We then prove that $\mathrm{MBR}(C_k, \{X, Z\}) = \mathrm{MBR}(C_k, \{X\})$.

Concerning the computation of $\mathrm{MBR}(C_k, \{X\})$, the $(C, \{X\})$ Bayes classifier assigns x, $x \in \mathcal{X}$, to 1 if and only if, recall (27),

$$\frac{f_{X|X+kY<0}(x)}{f_{X|X+kY\geq 0}(x)} \leq 1.$$

The required densities $f_{X|X+kY\geq 0}$ and $f_{X|X+kY<0}$ are known from Appendix Appendix A.2. We take this into account to re-write the expression above as

$$\frac{2\exp\left\{\frac{-x^2}{2\sigma_X^2\sigma_{X+kY}^2}\right\}\frac{1}{\sqrt{2\pi}\sigma_X}\Phi\left(-\frac{\rho\sqrt{\sigma_{X+kY}^2}x}{\sqrt{\sigma_X^2}}\frac{1}{\sqrt{\sigma_{X+kY}^2(1-\rho^2)}}\right)}{2\exp\left\{\frac{-x^2}{2\sigma_X^2\sigma_{X+kY}^2}\right\}\frac{1}{\sqrt{2\pi}\sigma_X}\Phi\left(\frac{\rho\sqrt{\sigma_{X+kY}^2}x}{\sqrt{\sigma_X^2}}\frac{1}{\sqrt{\sigma_{X+kY}^2(1-\rho^2)}}\right)} \leq 1,$$

where $\sigma_X^2 = 1$ is the variance of X, $\sigma_{X+kY}^2 = 1 + k^2$ is the variance of $X + kY$, and $\rho = \frac{1}{\sqrt{1+k^2}}$ is the correlation between X and $X + kY$.

Many terms cancel out, and we get simply

$$\Phi\left(-\frac{x}{k}\right) \leq \Phi\left(\frac{x}{k}\right).$$

As Φ is a non-decreasing function, this condition is the same as

$$-\frac{x}{k} \leq \frac{x}{k}.$$

After further cancellations of the denominators, we simply obtain the condition $-x \leq x$, which holds if and only if $x \geq 0$. As a result, the classifier assigns x to 1 if x is non-negative, and to 0 otherwise; note that the classifier applied to X gives X_{disc}.

Therefore, $\mathrm{MBR}(C_k, \{X\})$ depends on the angle, in the two-dimensional space defined by the pair (X, Y), between the lines associated with X and $X + kY$, $\arctan k$. Note that the only knowledge of X needed concerns the

ACCEPTED MANUSCRIPT

value that X_{disc} takes; recall (29). As a result, Figure A.5 also illustrates the regions where a wrong classification will occur, which are the regions (a) and (b). The probabilities associated with the different regions have been given in (A.3), allowing us to obtain

$$\text{MBR}(C_k, \{X\}) = 2\frac{\arctan k}{2\pi} = \frac{\arctan k}{\pi}.$$

We now verify that $\text{MBR}(C_k, \{X, Z\}) = \text{MBR}(C_k, \{X\})$. By (27), the $(C_k, \{X, Z\})$ Bayes classifier associated with the MBR takes the value 1 if and only if

$$\frac{f_{(X,Z)|X+kY<0}(x, z)}{f_{(X,Z)|X+kY\geq 0}(x, z)} \leq 1.$$

Using the facts that Z and X are class-conditionally independent and that Z is independent of C_k, the condition above reduces to

$$\frac{f_{X|X+kY<0}(x)}{f_{X|X+kY\geq 0}(x)} \leq 1.$$

As a result, the points (x, z) assigned by the classifier to the value 1 are those that verify $x \geq 0$. As a result, $\text{MBR}(C_k, \{X, Z\}) = \text{MBR}(C_k, \{X\})$.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

References

[1] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Recent advances and emerging challenges of feature selection in the context of big data, Knowledge-Based Systems 86 (2015) 33–45.

[2] E. P. Xing, M. I. Jordan, R. M. Karp, et al., Feature selection for high-dimensional genomic microarray data, in: ICML, Vol. 1, Citeseer, 2001, pp. 601–608.

[3] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, bioinformatics 23 (19) (2007) 2507–2517.

[4] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, A review of feature selection methods on synthetic data, Knowledge and information systems 34 (3) (2013) 483–519.

[5] T. Li, C. Zhang, M. Ogihara, A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, Bioinformatics 20 (15) (2004) 2429–2437.

[6] H. Liu, J. Li, L. Wong, A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns, Genome informatics 13 (2002) 51–60.

[7] F. Bagherzadeh-Khiabani, A. Ramezankhani, F. Azizi, F. Hadaegh, E. W. Steyerberg, D. Khalili, A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results, Journal of clinical epidemiology.

[8] Y. Yang, J. O. Pedersen, A comparative study on feature selection in text categorization, in: ICML, Vol. 97, 1997, pp. 412–420.

[9] M. Rogati, Y. Yang, High-performing feature selection for text classification, in: Proceedings of the eleventh international conference on Information and knowledge management, ACM, 2002, pp. 659–661.

ACCEPTED MANUSCRIPT

[10] P. Varela, A. Martins, P. Aguiar, M. Figueiredo, An empirical study of feature selection for sentiment analysis, in: 9th conference on telecommunications. Conftele 2013, Castelo Branco, 2013.

[11] F. H. Khan, U. Qamar, S. Bashir, Swims: Semi-supervised subjective feature weighting and intelligent model selection for sentiment analysis, Knowledge-Based Systems 100 (2016) 97–111.

[12] R. E. Schapire, Y. Singer, Boostexter: A boosting-based system for text categorization, Machine learning 39 (2) (2000) 135–168.

[13] K. Crammer, Y. Singer, A new family of online algorithms for category ranking, in: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2002, pp. 151–158.

[14] C. Pascoal, M. R. de Oliveira, R. Valadas, P. Filzmoser, P. Salvador, A. Pacheco, Robust feature selection and robust pca for internet traffic anomaly detection, in: INFOCOM, 2012 Proceedings IEEE, IEEE, 2012, pp. 1755–1763.

[15] J. Cai, J. Luo, S. Wang, S. Yang, Feature selection in machine learning: A new perspective, Neurocomputing 300 (2018) 70–79. doi:10.1016/j.neucom.2017.11.077.
URL https://doi.org/10.1016/j.neucom.2017.11.077

[16] R. Kohavi, G. H. John, Wrappers for feature subset selection, Artificial intelligence 97 (1) (1997) 273–324.

[17] I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh, Feature extraction: foundations and applications, Vol. 207, Springer, 2008.

[18] G. Brown, A. Pocock, M.-J. Zhao, M. Luján, Conditional likelihood maximisation: A unifying framework for information theoretic feature selection, J. Mach. Learn. Res. 13 (2012) 27–66.
URL http://dl.acm.org/citation.cfm?id=2188385.2188387

ACCEPTED MANUSCRIPT

[19] J. R. Vergara, P. A. Estévez, A review of feature selection methods based on mutual information, Neural Computing and Applications 24 (1) (2014) 175–186.

[20] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks 5 (1994) 537–550.

[21] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 1226–1238.

[22] J.-J. Huang, N. Lv, S.-Q. Li, Y.-Z. Cai, Feature selection for classificatory analysis based on information-theoretic criteria, Acta Automat. Sinica 34 (3) (2008) 383–392. doi:10.3724/SP.J.1004.2008.00381.
URL http://dx.doi.org/10.3724/SP.J.1004.2008.00381

[23] C. Pascoal, Contributions to variable selection and robust anomaly detection in telecommunications, Ph.D. thesis, Instituto Superior Técnico (2014).

[24] D. Lin, X. Tang, Conditional infomax learning: An integrated framework for feature extraction and fusion., in: A. Leonardis, H. Bischof, A. Pinz (Eds.), ECCV (1), Vol. 3951 of Lecture Notes in Computer Science, Springer, 2006, pp. 68–82.
URL http://dblp.uni-trier.de/db/conf/eccv/eccv2006-1.html#LinT06

[25] H. H. Yang, J. Moody, Data visualization and feature selection: New algorithms for nongaussian data, in: in Advances in Neural Information Processing Systems, MIT Press, 1999, pp. 687–693.

[26] F. Fleuret, Fast binary feature selection with conditional mutual information, The Journal of Machine Learning Research 5 (2004) 1531–1555.

ACCEPTED MANUSCRIPT

[27] C. Pascoal, M. R. Oliveira, A. Pacheco, R. Valadas, Theoretical evaluation of feature selection methods based on mutual information, Neurocomput. 226 (C) (2017) 168–181. doi:10.1016/j.neucom.2016.11.047.
URL https://doi.org/10.1016/j.neucom.2016.11.047

[28] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
URL http://dl.acm.org/citation.cfm?id=944919.944968

[29] C. E. Shannon, A mathematical theory of communication, Bell System Tech. J. 27 (1948) 379–423, 623–656.

[30] T. M. Cover, J. A. Thomas, Elements of information theory, 2nd Edition, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006.

[31] P. E. Meyer, G. Bontempi, On the use of variable complementarity for feature selection in cancer classification, in: Applications of Evolutionary Computing, Springer, 2006, pp. 91–102.

[32] S. Watanabe, Information theoretical analysis of multivariate correlation, IBM Journal of research and development 4 (1) (1960) 66–82.

[33] A. J. Bell, The Co-Information Lattice, in: ICA 2003, Nara, Japan, 2003.

[34] P. E. Meyer, C. Schretter, G. Bontempi, Information-theoretic feature selection in microarray data using variable complementarity, Selected Topics in Signal Processing, IEEE Journal of 2 (3) (2008) 261–274.

[35] N. X. Vinh, S. Zhou, J. Chan, J. Bailey, Can high-order dependencies improve mutual information based feature selection?, Pattern Recognition.

[36] H. Cheng, Z. Qin, C. Feng, Y. Wang, F. Li, Conditional mutual information-based feature selection analyzing for synergy and redundancy, ETRI Journal 33 (2) (2011) 210–218.

ACCEPTED MANUSCRIPT

[37] 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.

[38] M. Bennasar, Y. Hicks, R. Setchi, Feature selection using joint mutual information maximisation, Expert Systems with Applications 42 (22) (2015) 8520–8532.

[39] N. Kwak, C.-H. Choi, Input feature selection for classification problems, Neural Networks, IEEE Transactions on 13 (1) (2002) 143–159. doi:10.1109/72.977291.
URL http://dx.doi.org/10.1109/72.977291

[40] N. Hoque, D. Bhattacharyya, J. K. Kalita, Mifs-nd: a mutual information-based feature selection method, Expert Systems with Applications 41 (14) (2014) 6371–6385.

[41] A. Jakulin, Machine learning based on attribute interactions, Ph.D. thesis, Univerza v Ljubljani (2005).

[42] A. El Akadi, A. El Ouardighi, D. Aboutajdine, A powerful feature selection approach based on mutual information, International Journal of Computer Science and Network Security 8 (4) (2008) 116.

[43] J. R. Vergara, P. A. Estévez, CMIM-2: an enhanced conditional mutual information maximization criterion for feature selection, Journal of Applied Computer Science Methods 2.

[44] P. A. Estévez, M. Tesmer, C. A. Perez, J. M. Zurada, Normalized mutual information feature selection, Neural Networks, IEEE Transactions on 20 (2) (2009) 189–201.

[45] R. A. Johnson, D. W. Wichern, Applied Multivariate Statistical Analysis (6th Edition), Pearson.

ACCEPTED MANUSCRIPT

[46] S. Kumar, W. Byrne, Minimum bayes-risk decoding for statistical machine translation, Tech. rep., DTIC Document (2004).

[47] V. Goel, W. J. Byrne, Minimum bayes-risk automatic speech recognition, Computer Speech & Language 14 (2) (2000) 115–135.

[48] A. Azzalini, Further results on a class of distributions which includes the normal ones, Statistica (Bologna) 46 (2) (1986) 199–208.

[49] A. F. Karr, Probability, Springer Texts in Statistics, Springer-Verlag, New York, 1993. doi:10.1007/978-1-4612-0891-4.
URL http://dx.doi.org/10.1007/978-1-4612-0891-4

[50] K. Dadkhah, H. Midi, O. S. Sharipov, The performance of mutual information for mixture of bivariate normal distributions based on robust kernel estimation, Applied Mathematical Sciences 4 (29) (2010) 1417–1436.

[51] M. Abramowitz, I. A. Stegun, Handbook of mathematical functions: with formulas, graphs, and mathematical tables, Vol. 55, Courier Corporation, 1964.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT



Francisco Macedo graduated in Applied Mathematics and Computer Science at IST, (Instituto Superior Técnico, Universidade de Lisboa, Portugal) on 2011, received the M.Sc. degree in Mathematics and Applications from IST on 2012, and obtained, the Joint PhD in Computational and Stochastic Mathematics from Instituto Superior Técnico and École Polytechnique Fédérale de Lausanne (EPFL) on 2016. He was a teaching assistant at IST and EPFL. His areas of interest are related with sports quantitative analysis, portfolio management, data science, stochastic processes, and numerical analysis.



M. Rosário Oliveira graduated in applied mathematics and computer science at Instituto Superior Técnico, in 1992, received the M.Sc. degree in Applied Mathematics from Universidade Técnica de Lisboa in 1995, and obtained the Ph.D. degree also in Mathematics at the Universidade Técnica de Lisboa in 2002. She has been working at Instituto Superior Técnico (IST) since 1990, where she is Assistant Professor at the Mathematics Department, and has been a full member of the Center for Mathematics and its Applications from IST since 2002. She has been involved in teaching, research and consulting in the areas of data science and statistics, and in national and international projects/consortiums, including European Networks of Excellence Euro-NGI, Euro-FGI, and TEMPUS project Applied Computing in Engineering and Science. Her current research interests include data science, robust statistics, multivariate analysis, and biostatistics.

ACCEPTED MANUSCRIPT



António Pacheco is Full Professor in Probability and Statistics at Instituto Superior Técnico (IST), Universidade de Lisboa, Portugal. He obtained a Degree in Probability and Statistics from the Universidade de Lisboa in 1987, a MSc degree in Applied Mathematics from the Universidade Técnica de Lisboa in 1991, a PhD degree in Operations Research from Cornell University in 1994, and Habilitation in Mathematics from IST in 2006. He has been involved in teaching, research and consulting in the areas of statistics and applied stochastic processes, and in national and international projects/consortiums, including the European Networks of Excellence Euro-NGI and Euro-FGI. He has published more than 100 research works, including 3 books and around 50 papers in international peer-reviewed journals.



Rui Valadas is a Full Professor of the Dept. of Electrical and Computer Engineering at Instituto Superior Técnico (IST), Universidade de Lisboa, and a Senior Researcher at Instituto de Telecomunicações (IT). He is the director of the Computer Networking BSc of IST. He has been the team leader of IT in the Euro-NGI, Euro-FGI, and Euro-NF Networks of Excellence, funded by the European Commission. Over the years, he has published more 100 papers in the areas Optical Wireless Communications, Network Dimensioning, and Internet Traffic Modeling and Statistical Characterization.

ACCEPTED MANUSCRIPT