Neural Networks 48 (2013) 1–7

Contents lists available at ScienceDirect

Neural Networks

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

Neural networks letter

Is mutual information adequate for feature selection in regression?

Benoît Frénay [*,1], Gauthier Doquire [1], Michel Verleysen

Machine Learning Group - ICTEAM, Université catholique de Louvain, Place du Levant 3, 1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Article history:
Received 17 September 2012
Revised and accepted 4 July 2013

Keywords:
Mutual information
Feature selection
Regression
MSE
MAE

ABSTRACT

Feature selection is an important preprocessing step for many high-dimensional regression problems. One of the most common strategies is to select a relevant feature subset based on the mutual information criterion. However, no connection has been established yet between the use of mutual information and a regression error criterion in the machine learning literature. This is obviously an important lack, since minimising such a criterion is eventually the objective one is interested in. This paper demonstrates that under some reasonable assumptions, features selected with the mutual information criterion are the ones minimising the mean squared error and the mean absolute error. On the contrary, it is also shown that the mutual information criterion can fail in selecting optimal features in some situations that we characterise. The theoretical developments presented in this work are expected to lead in practice to a critical and efficient use of the mutual information for feature selection.

1. Introduction

In many regression problems, input data are originally high-dimensional. As an example, in the field of near-infrared spectroscopy analysis, each sample is described by tens or hundreds of features, corresponding to its spectrum components. Much of these features are in practice either redundant or irrelevant to the considered regression problem (Rossi, Lendasse, Francois, Wertz, & Verleysen, 2006).

It is well known that learning with a huge number of features and limited sample size is a hard task because of the so-called *curse of dimensionality* (Bellman, 1961) and its consequences (Verleysen, 2003). In such settings, the risk of overfitting is high, especially when complex models (with numerous parameters) have to be inferred from the data.

In order to address the aforementioned issues, one of the most popular methods is to perform feature selection before any further learning step. The idea is to select a small subset of features which are together highly relevant with the output to predict (see Guyon & Elisseeff, 2003 for a nice introduction). Feature selection has the advantage over projection methods (which project the features onto a space of small dimension) that the original features are not transformed, which allows one to subsequently build easy-to-interpret models.

The feature selection problem should be distinguished from the one of sufficient dimension reduction (Globerson & Tishby, 2003), where the objective is to obtain a subspace of minimal dimension containing the whole information about the output. On the contrary, the goal of feature selection is to select only a few of the original features, even if the price to pay is a small loss of information. Feature selection allows one to considerably reduce the dimension of the dataset, which can improve the performance of the prediction model by reducing the effects of the curse of dimensionality. It thus also speeds up the learning process and leads to a better understanding of the considered problem.

An intuitive and appealing idea is to select the features based on the performance of an inference model. This approach, called *wrapper* in the literature (Kohavi & John, 1997), often leads to good prediction performances but also suffers from two main drawbacks. First, it can be very computationally demanding since many prediction models with different feature subsets have to be built. Then, the results of the wrapper strategy lack generality as their use is limited to a specific regression model. To circumvent both problems, filter methods are often used in practice. Such methods are based on a relevance criterion measuring the quality of feature subsets; this criterion is independent of any prediction algorithm. Filters are traditionally much faster than wrappers and can be used with any regression algorithm. Among the numerous solutions proposed in the literature, mutual information (Shannon, 1948) is one of the most popular relevance criteria, due to many advantages for the feature selection task which will be detailed in the next section. It has therefore been used in a large number of works (see e.g. Dijck & Hulle, 2006, Fleuret, 2004, Kojadinovic & Wottka, 2000, Rossi, François, Wertz, Meurens, & Verleysen, 2007) since the seminal paper Battiti (1994).

* Corresponding author. Tel.: +32 10 47 81 33; fax: +32 10 47 25 98.
E-mail addresses: benoit.frenay@uclouvain.be (B. Frénay), gauthier.doquire@uclouvain.be (G. Doquire), michel.verleysen@uclouvain.be (M. Verleysen).

[1] Both authors contributed equally to this work.

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.07.003

The eventual objective in a regression problem is to reduce as much as possible an error criterion; the most frequently used ones are the mean squared error (MSE) and the mean absolute error (MAE). However, to the best of our knowledge, no explicit connection has been established in the machine learning literature between the use of the mutual information as a feature selection criterion and the MSE or the MAE. In information theory, the MSE has been e.g. related to the derivative of mutual information with respect to the signal to noise ratio (Guo, Shamai, & Verdú, 2005). Moreover, there exists a relationship between MSE and mutual information (see Section 3.2 and e.g. Chen, Hu, Li, & Sun, 2008, Ihara, 1993), but this relation is only valid in the case of Gaussian estimation errors. This paper addresses the previously discussed lack of connection in machine learning by showing that, assuming some realistic hypotheses on the estimation errors on the target, the mutual information criterion is actually optimal from a MSE or a MAE point of view. This result confirms that the mutual information is a criterion which is worth considering for feature selection. In addition, the paper also illustrates the fact that in some situations, the features selected using the mutual information criterion are not the ones minimising the considered error criterion. In such cases, mutual information should not necessarily be the criterion of choice, and the results of the feature selection procedure should be carefully analysed. This work thus intends to present both theoretical arguments and case studies of the potential interest of mutual information for feature selection in regression problems; the final goal is to better apprehend its behaviour in order to use it in the most efficient and sensitive way. It should be noted that the results in this paper assume the knowledge of the distributions of the random variables, and can thus be seen as infinite-sample arguments. The variance of the mutual information, MSE and MAE estimators are not considered here. A preliminary study on the adequation between mutual information and misclassification probability for classification problems was published in Frénay, Doquire, and Verleysen (2012, 2013). This paper extends the approach and focus on regression tasks.

The remaining of the paper is organised as follows. Section 2 recalls basic notions about mutual information and entropy for feature selection. The well known MSE and MAE error criteria are presented and linked to the estimation error on the target output in the context of feature selection. Section 3 theoretically demonstrates the optimality of the mutual information for three popular models of the estimation error. Section 4, on the contrary, illustrates the potential inadequacy of mutual information and characterises the situations where this criterion is likely to fail. Section 5 briefly discusses the results while Section 6 concludes the work.

2. Theory and notations

This section briefly gives basic definitions about mutual information and entropy. The MSE and the MAE, the two error criteria considered in this work, are then briefly reviewed. Next, Sections 3 and 4 will establish a connection between the use of mutual information for feature selection and the two error criteria.

2.1. Mutual information and entropy

Mutual information (Shannon, 1948) is a quantity measuring the dependency existing between two (groups of) random variables, assumed to be continuous in this work. Let X and Y be random variables whose respective probability density functions are f_X and f_Y and whose domains are $\mathcal{X}$ and $\mathcal{Y}$. Let us also define the joint probability density function $f_{X,Y}$. The mutual information between X and Y is defined as

$$I(X;Y) = -\int_{\mathcal{X}}\int_{\mathcal{Y}} f_{X,Y}(x,y)\ \log\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\,dxdy. \tag{1}$$

Eq. (1) can actually be rewritten in terms of entropy and conditional entropy, respectively defined as

$$H(X) = -\int_{\mathcal{X}} f_X(x)\ \log f_X(x)\,dx \tag{2}$$

and

$$H(Y|X) = -\int_{\mathcal{X}}\int_{\mathcal{Y}} f_{X,Y}(x,y)\ \log\frac{f_X(x)}{f_{X,Y}(x,y)}\,dxdy. \tag{3}$$

Using Eqs. (1)–(3), it is possible to write

$$I(X;Y) = H(Y) - H(Y|X). \tag{4}$$

The mutual information can thus be understood as the reduction of uncertainty (measured by the entropy) on the values of Y once X is known. If Y denotes the target output to predict and X a subset of features, mutual information has a quite natural interpretation as a feature selection criterion. Indeed, a feature subset having a high mutual information with the target output is likely to reduce the uncertainty on the values taken by the output, what is obviously desirable. Besides an intuitive interpretation, mutual information also has the advantage of detecting non-linear relationships between variables while some other popular criteria (such as the correlation coefficient) are essentially limited to linear dependencies. Eventually, mutual information can naturally be defined for groups of variables, making it possible to evaluate subsets of features; this last property is of crucial importance when jointly redundant or relevant features make univariate criteria useless.

For a given regression problem between inputs X and output Y, $H(Y)$ is fixed and does not depend on the choice of features. Therefore, from a feature selection point of view, Eq. (4) indicates that selecting features X in order to maximise $I(X;Y)$ can be achieved by selecting features which minimise $H(Y|X)$. In the remainder of the paper, the discussion will be about $H(Y|X)$, while the same conclusions can be drawn about $I(X;Y)$.

2.2. Regression error criteria

As mentioned in Section 1, the final objective in a regression problem is to minimise an error criterion, measuring in some way the difference between the predicted value and the actual value of the output. In this section, two popular error criteria are reviewed and linked to the estimation error on the target output.

Let us assume that for a given subset of d features, the output $Y \in \mathfrak{R}$ depends probabilistically on the input $X \in \mathfrak{R}^d$. Moreover, the function f provides an estimate $\hat{Y} = f(X)$ of Y given X. Then, the estimation error is

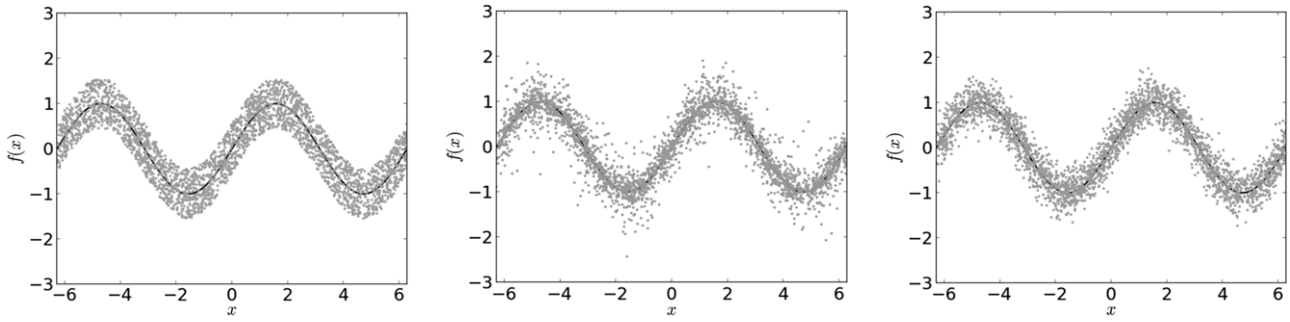
$$\epsilon = f(X) - Y, \tag{5}$$

whose zero-mean distribution depends on the choice of features. Two popular regression error criteria can be rewritten in terms of ϵ, which are discussed below.

The mean square error (MSE) of the estimate f is defined as the variance $\mathrm{E}\left\{(f(X)-Y)^2\right\} = \mathrm{E}\left\{\epsilon^2\right\}$ of ϵ, where $\mathrm{E}\{\cdot\}$ denotes the expected value. Another popular error criterion is the mean absolute error (MAE), defined as the expected absolute value $\mathrm{E}\{|f(X)-Y|\} = \mathrm{E}\{|\epsilon|\}$ of ϵ. In feature selection, one is typically interested in feature subsets which allow one to obtain estimates achieving low MSE or MAE values.

2.3. Error criteria and entropy of estimation error

For a given estimate f, it is well known (see e.g. Ash, 1990, Cover & Thomas, 1991) that the conditional entropy of Y given X can be



(a) Uniform noise $\mathcal{U}(\mu = 0, \Delta = 1.105)$. (b) Laplacian noise $\mathcal{L}(\mu = 0, \lambda = 0.203)$. (c) Gaussian noise $\mathcal{N}(\mu = 0, \sigma = 0.267)$.

Fig. 1. Functional $f(x) = \sin(x)$ polluted by uniform, Laplacian or Gaussian target noise with identical conditional target entropy $H(Y|X) = 0.1$.

rewritten in terms of ϵ as

$$H(Y|X) = H(\epsilon|X). \tag{6}$$

To show this relationship, let us rewrite $H(Y|X)$ as

$$\begin{aligned} H(Y|X) &= \int_{\mathcal{X}} f_X(x) H(Y|X = x) dx \\ &= \int_{\mathcal{X}} f_X(x) H(f(X) + \epsilon|X = x) dx. \end{aligned} \tag{7}$$

Since $f(X)$ is fixed when X is known and since the differential entropy is translation invariant ($H(X + k) = H(X)$ for a constant k, see e.g. Emmert-Streib & Dehmer, 2008), Eq. (6) follows directly from Eq. (7).

The rest of this paper considers the relationship between the mutual information and the two error criteria in different settings. Assuming f and the entropy or mutual information estimator can be accurately estimated using the available data (see e.g. Kozachenko & Leonenko, 1987, Kraskov, Stögbauer, & Grassberger, 2004 for mutual information estimation), the results obtained in this paper show the interest of using mutual information as a feature selection criterion for real-world problems.

3. Mutual information adequacies

This section shows how mutual information can be an adequate criterion for feature selection in regression. More specifically, when the conditional distribution of the estimation error is uniform, Laplacian or Gaussian, choosing the feature subset which minimises the conditional target entropy $H(Y|X)$ is equivalent to minimising either the MSE or the MAE criterion. Here, X corresponds to a specific subset of features and $H(Y|X)$ depends on the choice of this subset X. Moreover, it is assumed that when two feature subsets are compared for a given dataset, the distributions of the estimation error belong to the same parametric family (uniform, Laplacian or Gaussian) in both cases. This hypothesis is realistic when feature subsets which are not too different (in terms of informative contain) are compared, like e.g. at a given step of a forward or backward search.

3.1. Specification of regression examples

As explained in Section 2, mutual information is adequate for feature selection in regression with respect to an error criterion if minimising the conditional target entropy $H(Y|X)$ always improves the criterion. The choice of the criterion depends on the application, so it could also be true for mutual information adequacy.

In this section, three realistic estimation error distributions are considered: a uniform, a Laplacian and a Gaussian distribution. The estimation error is assumed to be identically distributed for any $x \in \mathcal{X}$, which means that the conditional entropy $H(Y|X)$ is equal to the specific conditional entropy $H(Y|X = x)$ for any $x \in \mathcal{X}$. Since the means of the estimation error distributions are zero, they only have one effective parameter: the width Δ for the uniform estimation error, the scale λ for the Laplacian estimation error and the standard deviation σ for the Gaussian estimation error. The value of the estimation error distribution parameter (Δ, λ or σ) depends on the feature subset which corresponds to X. Considering these estimation error distributions, the conditional target entropies $H(Y|X)$ are $\ln \Delta$, $\ln[2e\lambda]$ and $\frac{1}{2}\ln\left[2\pi e\sigma^2\right]$, respectively (Cover & Thomas, 1991).

In order to visualise each type of estimation error, Fig. 1 shows an example of functional $f(x) = \sin(x)$ which is polluted by three types of noise, with identical conditional target entropy $H(Y|X) = 0.1$. Under uniform noise, the function values stay inside a tube around the real function f. Under Laplacian or Gaussian noise, the distribution of function values spreads around the real function f, with more or less thick tails.

3.2. Adequacy assessment for the MSE criterion

As shown in Section 2, the MSE can be interpreted as the expected variance of the estimation error. Since the estimation error is assumed to be identically distributed for any $x \in \mathcal{X}$, its variance is precisely equal to the MSE. For the uniform, Laplacian and Gaussian estimation errors, the variance can be written in terms of their only free parameter as $\frac{\Delta^2}{12}$, $2\lambda^2$ and σ^2, respectively (Cover & Thomas, 1991; Kotz, Kozubowski, & Podgórski, 2001). Using the expressions for the conditional target entropies and these relationships, it is possible to express the MSE in sole terms of $H(Y|X)$. For the uniform, Laplacian and Gaussian estimation error, the MSE becomes $\frac{1}{12}\exp[2H(Y|X)]$, $\frac{1}{2e^2}\exp[2H(Y|X)]$ and $\frac{1}{2\pi e}\exp[2H(Y|X)]$, respectively. Notice that the MSE no longer depends explicitly on the parameter of the estimation error distribution. In the Gaussian case, similar relationships between MSE and conditional entropy have been reported e.g. in Chen et al. (2008), Guo et al. (2005), Ihara (1993).

In the above relationships, the MSE is a monotonically increasing function of the conditional target entropy, which depends on the selected feature subset. It means that if different feature subsets are compared and if the distribution of the estimation errors belongs to the parametric family (uniform, Laplacian or Gaussian) in each case, choosing the feature subset which minimises the conditional target entropy $H(Y|X)$ necessarily corresponds to minimising the MSE. This is illustrated in Fig. 2 which shows the MSE in terms of the conditional target entropy $H(Y|X)$ for the uniform, Laplacian and Gaussian estimation errors. Since the Gaussian distribution is the maximum entropy distribution for a given estimation error variance σ^2 (Cover & Thomas, 1991), the curve corresponding to the Gaussian error gives a lower bound for the MSE and defines an admissible region for the MSE as shown in Fig. 2.

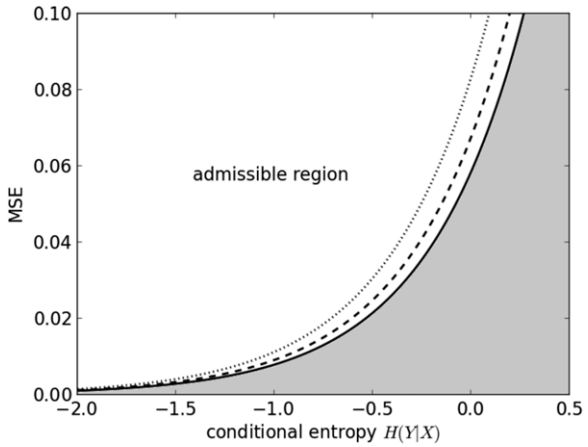


Fig. 2. MSE in terms of the conditional target entropy $H(Y|X)$ for identically distributed uniform (dotted line), Laplacian (dashed line) or Gaussian (plain line) estimation error. The Gaussian curve gives a lower bound and defines an admissible region (in white).

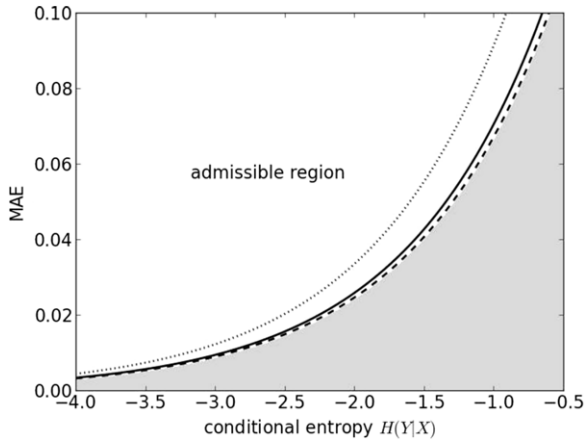


Fig. 3. MAE in terms of the conditional target entropy $H(Y|X)$ for identically distributed uniform (dotted line), Laplacian (dashed line) or Gaussian (plain line) estimation error. The Laplacian curve gives a lower bound and defines an admissible region (in white).

3.3. Adequacy assessment for the MAE criterion

Since the estimation error is assumed to be identically distributed, the MAE is by definition equal to the expected absolute value of the estimation error for any $x \in \mathcal{X}$. For the uniform, Laplacian and Gaussian estimation errors, the expected absolute value can be written in terms of their only free parameter as $\frac{A}{4}$, λ and $\sqrt{\frac{2}{\pi}}\sigma$, respectively (Kotz et al., 2001). Using the expressions for the conditional target entropies and these relationships, it is possible to express the MAE in sole terms of $H(Y|X)$. For the uniform, Laplacian and Gaussian estimation error, the MAE becomes $\frac{1}{4}\exp\left[H(Y|X)\right]$, $\frac{1}{2e}\exp\left[H(Y|X)\right]$ and $\frac{1}{\pi\sqrt{e}}\exp\left[H(Y|X)\right]$, respectively. Notice that the MAE no longer depends explicitly on the estimation error distribution parameter.

In the above relationships, the MAE is a monotonically increasing function of the sole conditional target entropy, which depends on the selected feature subset. It means that if different feature subsets are compared and if the distribution of the estimation errors belongs to the parametric family (uniform, Laplacian or Gaussian) in each case, choosing the feature subset which minimises the conditional target entropy $H(Y|X)$ necessarily corresponds to minimising the MAE. This is illustrated in Fig. 3 which shows the MAE in terms of the conditional target entropy $H(Y|X)$ for the uniform, Laplacian and Gaussian estimation errors. Since the Laplacian distribution is the maximum entropy distribution for a given expected estimation error absolute value λ (Kotz et al., 2001), the corresponding curve gives a lower bound for the MAE and defines an admissible region.

3.4. Short discussion

This section shows that mutual information is an adequate criterion for feature selection with respect to the MSE and the MAE, when the estimation error is identically distributed for any $x \in \mathcal{X}$ with a uniform, Laplacian or Gaussian distribution. Indeed, maximising mutual information is equivalent to minimising the conditional target entropy $H(Y|X)$, which in the above settings also corresponds to minimising either the MSE or the MAE.

4. Mutual information inadequacies

This section shows that mutual information is not always adequate for feature selection in regression. In the proposed example, the conditional distribution of the estimation error is assumed to be a Student distribution. For this particular setting, it is shown that choosing the feature subset which minimises the conditional target entropy $H(Y|X)$ is not necessarily equivalent to minimising either the MSE or the MAE criterion.

4.1. Specification of regression examples

In this section, it is assumed that the estimation error follows an identical Student distribution for any $x \in \mathcal{X}$. This distribution is often used for the robust modelling of random variables which look Gaussian, but whose distribution has ticker tails (Archambeau, Delannay, & Verleysen, 2008; Peel & McLachlan, 2000). The density of the non-standardised Student distribution is

$$\mathcal{S}\left(\epsilon = e|\mu, \nu, \sigma\right) = \frac{1}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)\sqrt{\nu\sigma^2}}\left[1 + \frac{(e-\mu)^2}{\nu\sigma^2}\right]^{-\frac{\nu+1}{2}} \tag{8}$$

where B is the beta function. Since the mean of the estimation error is zero, so is the parameter μ and the Student distribution only has two effective parameters: the number of degrees of freedom ν and the scale σ. The value of the distribution parameters ν and σ depend on the selected feature subset X. The conditional target entropy $H(Y|X)$ for a Student estimation error is

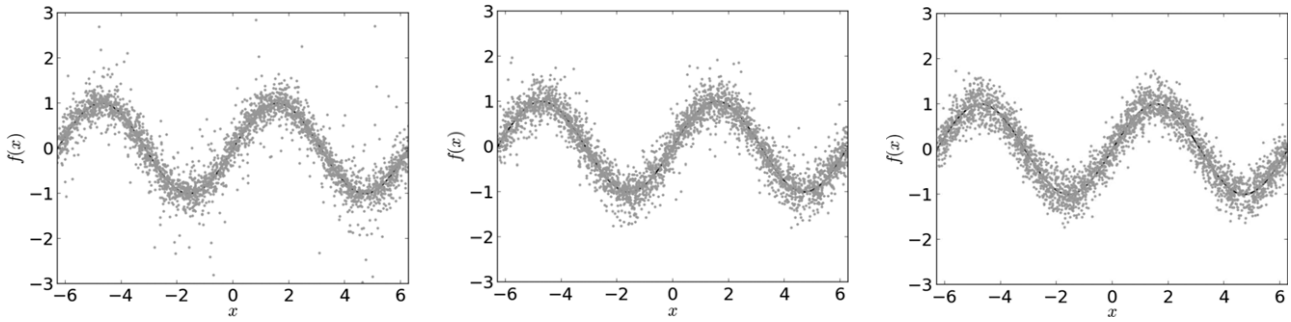
$$\left(\frac{\nu+1}{2}\right)\left[\Psi\left(\frac{\nu+1}{2}\right) - \Psi\left(\frac{\nu}{2}\right)\right] + \ln\left[\sqrt{\nu}B\left(\frac{1}{2}, \frac{\nu}{2}\right)\right] + \ln\sigma \tag{9}$$

where Ψ is the digamma function (Cover & Thomas, 1991). Fig. 4 shows an example of functional $f(x) = \sin(x)$ which is polluted by Student noises with different numbers of degrees of freedom but identical conditional target entropy $H(Y|X) = 0.1$. The spread of the distribution and the thickness of its tail depend on ν.

4.2. Inadequacy assessment for the MSE criterion

For a Student estimation error which is identical for any $x \in \mathcal{X}$, the variance can be expressed in terms of its two free parameters as $\frac{\nu}{\nu-2}\sigma^2$. Hence, using this relationship and the expression of the conditional target entropy, the MSE can be rewritten in terms of the number of degrees of freedom ν and $H(Y|X)$ as

$$\mathrm{E}\left\{(Y - f(X))^2\right\} = \frac{1}{(\nu-2)\,B\left(\frac{1}{2}, \frac{\nu}{2}\right)^2}\exp\left\{2H(Y|X) - (\nu+1)\left[\Psi\left(\frac{\nu+1}{2}\right) - \Psi\left(\frac{\nu}{2}\right)\right]\right\}. \tag{10}$$



(a) Student noise $\mathcal{S}(\mu = 0, \nu = 2.3, \sigma = 0.168)$. (b) Student noise $\mathcal{S}(\mu = 0, \nu = 5, \sigma = 0.217)$. (c) Student noise $\mathcal{S}(\mu = 0, \nu = 30, \sigma = 0.259)$.

Fig. 4. Functional $f(x) = \sin(x)$ polluted by Student noises with different parameters but identical conditional target entropy $H(Y|X) = 0.1$.

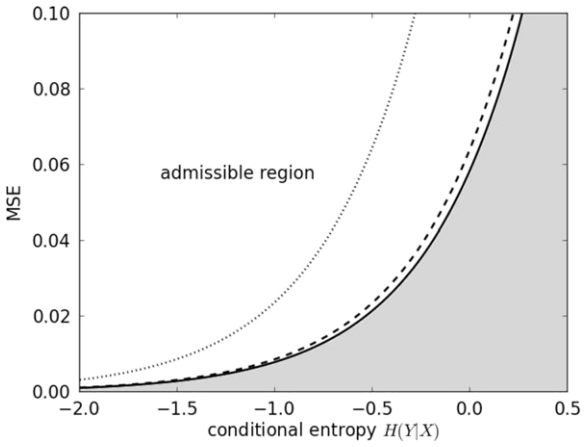


Fig. 5. MSE in terms of the conditional target entropy $H(Y|X)$ for Student estimation error with different numbers of degrees of freedom: $\nu = 2.3$ (dotted line), $\nu = 5$ (dashed line) or $\nu = 30$ (plain line). The admissible region defined by the Gaussian curve appears in white.

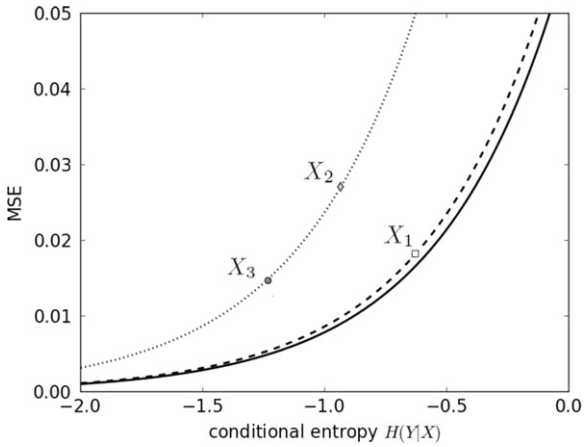


Fig. 6. Example of mutual information failure for Student estimation error with respect to the MSE. The candidate feature subsets correspond to different numbers of degrees of freedom: $\nu = 2.3$ (dotted line) and $\nu = 5$ (dashed line). The curve for $\nu = 30$ (plain line) is also shown for discussion (see text). The symbols X_i are feature subsets.

It shows that the MSE cannot be written in sole terms of the conditional target entropy. Indeed, the MSE still depends on the number of degrees of freedom ν (one could alternatively use σ, but ν is easier to understand intuitively). This is illustrated in Fig. 5 which shows the MSE in terms of the conditional target entropy $H(Y|X)$ for different numbers of degrees of freedom. Since the numbers of degrees of freedom of the Student estimation error distribution for different feature subsets are not necessarily identical (for example, data which are outliers with respect to some features can look normal with respect to other features), it is possible to decrease the conditional target entropy while increasing the MSE. This means that choosing the feature subset which minimises the conditional target entropy $H(Y|X)$ does not necessarily correspond to minimising the MSE.

Fig. 6 shows an example of mutual information failure with respect to the MSE, where two candidate features subsets X_1 and X_2 are characterised by a Student estimation error with parameters $\nu = 2.3$ and $\nu = 5$, respectively. Using mutual information, X_2 will be chosen rather than X_1, since $H(Y|X_2)$ is smaller than $H(Y|X_1)$. However, because of the different degrees of freedom of the estimation error affecting both feature subsets, the MSE is larger for X_2 than for X_1. Hence, selecting X_2 based on mutual information leads here to an increase in the criterion which should be minimised; mutual information fails as feature selection criterion. Notice that, in Fig. 6, X_3 is characterised by the same degree of freedom than X_2, but mutual information does not fail when choosing between X_1 and X_3. Indeed, selecting the feature subset X_3 because $H(Y|X_3)$ is smaller than $H(Y|X_1)$ effectively minimises the MSE.

In practice, the case of mutual information failures discussed in this section may be of little impact. Indeed, Fig. 6 also shows that the curve for $\nu = 30$ is quite close to the curve for $\nu = 5$. Hence, it is less dangerous to compare feature subsets with such estimation error distributions, which are common in practice. Since $\nu = 2.3$ is quite extreme as can be seen in Fig. 4, one can use mutual information in most practical cases with a quite reasonable confidence.

4.3. Inadequacy assessment for the MAE criterion

For the Student estimation error, the expected absolute value can be expressed in terms of its two free parameters (Psarakis & Panaretos, 1990) as

$$\mathrm{E}\{|Y - f(X)|\} = \frac{2\sqrt{\nu\sigma^2}}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)(\nu - 1)}. \tag{11}$$

Hence, in terms of the number of degrees of freedom ν and the conditional target entropy $H(Y|X)$, the MAE is

$$\mathrm{E}\{|Y - f(X)|\} = \frac{2}{(\nu - 1)\,B\left(\frac{1}{2}, \frac{\nu}{2}\right)^2} \exp\left\{H(Y|X) - \left(\frac{\nu + 1}{2}\right)\left[\Psi\left(\frac{\nu + 1}{2}\right) - \Psi\left(\frac{\nu}{2}\right)\right]\right\}. \tag{12}$$

As for the MSE, the MAE still depends on the number of degrees of freedom ν and cannot be written in sole terms of the conditional target entropy. This is illustrated in Fig. 7 which shows the MAE in

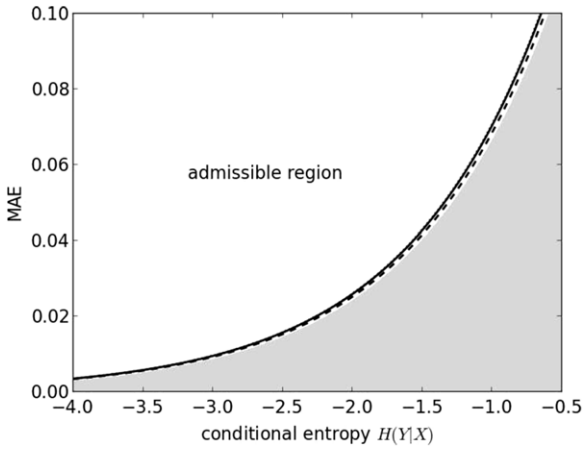


Fig. 7. MAE in terms of the conditional target entropy $H(Y|X)$ for Student estimation error with different numbers of degrees of freedom: $\nu = 2.3$ (dotted line), $\nu = 5$ (dashed line) or $\nu = 30$ (plain line). The admissible region defined by the Laplacian curve appears in white.

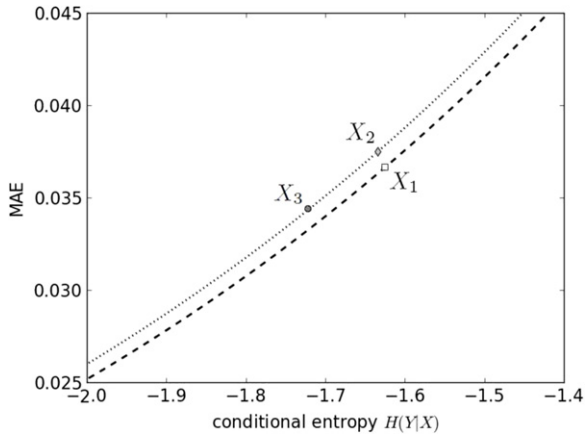


Fig. 8. Example of mutual information failure for Student estimation error with respect to the MAE. The candidate feature subsets correspond to different numbers of degrees of freedom: $\nu = 2.3$ (dotted line) and $\nu = 5$ (dashed line). The symbols X_i are feature subsets.

terms of the conditional target entropy $H(Y|X)$ for different numbers of degrees of freedom. The different curves are very close, yet they do not coincide and it is possible to decrease the conditional target entropy while increasing the MAE. Minimising the conditional target entropy $H(Y|X)$ does not necessarily correspond to minimising the MAE, but the problem is less important than for the MSE because the curves are very close from each other.

Fig. 8 shows an example of mutual information failure with respect to the MAE, where two candidate features subsets X_1 and X_2 are characterised by a Student estimation error with parameters $\nu = 2.3$ and $\nu = 5$, respectively. Using mutual information, X_2 will be chosen rather than X_1, since $H(Y|X_2)$ is smaller than $H(Y|X_1)$. However, selecting X_2 leads here to an increase in the MAE, which should rather be minimised. Fig. 8 also shows a counterexample: mutual information does not fail when choosing between the feature subsets X_1 and X_3, with two different Student estimation errors.

4.4. Short discussion

This section shows that mutual information is not always an adequate criterion to perform feature selection in regression. Indeed, when the estimation error has several parameters, it may be possible to obtain different values of the criterion for a given conditional target entropy $H(Y|X)$ (and *vice versa*). Intuitively, the knowledge of the value of the conditional target entropy $H(Y|X)$ only fixes one degree of freedom of the estimation error distribution parameters. Hence, problems may occur when the estimation error distribution is characterised by several parameters. In such a case, it becomes possible to select a feature subset which decreases the conditional target entropy with respect to other feature subsets while simultaneously increasing the MSE or the MAE. However, the impact of mutual information issues is likely to remain limited in terms of MSE and MAE for Student estimation errors. This is the case when the number of degrees of freedom is not too small, which is verified unless there is a large number of outliers.

5. Discussion

The examples in Section 3 show that mutual information is often a valuable criterion for feature selection in regression. Indeed, for realistic estimation errors with e.g. uniform, Laplacian or Gaussian distribution, choosing a feature subset which minimises the mutual information corresponds to minimising either the MSE or the MAE. Since it is often assumed (i) that the estimation error follows one of these distributions and (ii) that the MSE or the MAE is a sensible criterion, mutual information can in most cases be used safely. In fact, one can postulate that mutual information can be used whenever the estimation error is identically distributed for any $x \in \mathcal{X}$ and the estimation error distribution can be characterised by only one parameter.

Unfortunately, mutual information is not always optimal. Indeed, the example in Section 4 shows that when the estimation error distribution is characterised by multiple parameters like e.g. the Student distribution, it may be possible to obtain different values of the MSE or the MAE for a given value of the mutual information. Hence, minimising the mutual information does not necessarily correspond to minimising either the MSE or the MAE. However, it must be noticed that the impact of this issue may be of various importance. Indeed, in Section 4, the MSE can be quite different for a given conditional target entropy when the number of degrees of freedom of the Student distribution changes, whereas the difference remains quite small for the MAE. However, since it is not common to observe very small degrees of freedom (unless there is a large number of outliers), one can expect the impact of mutual information failures to remain small in practice, as discussed in Sections 4.2 and 4.4.

In Sections 3 and 4, the estimation error is assumed to be identically distributed for any $x \in \mathcal{X}$. However, this is not necessarily the case; the distribution of the estimation error may depend on x. For example, let us consider a simple estimation error which follows a Gaussian distribution $\mathcal{N}(0, \sigma_1)$ for one half of the samples and $\mathcal{N}(0, \sigma_2)$ for the other half. In terms of the standard deviations σ_1 and σ_2, the conditional target entropies $H(Y|X)$ for this non-identically distributed (n.i.d.) Gaussian estimation error is $\frac{1}{2}\ln[2\pi e\sigma_1\sigma_2]$, whereas the MSE is $\frac{1}{2}\left(\sigma_1^2 + \sigma_2^2\right)$ and the MAE is $\frac{1}{\sqrt{2\pi}}(\sigma_1 + \sigma_2)$. Here, it is possible to rewrite the MSE and the MAE by replacing e.g. σ_2, what gives

$$\frac{1}{2}\left(\sigma_1^2 + \frac{\exp[4H(Y|X)]}{4\pi^2 e^2 \sigma_1^2}\right) \tag{13}$$

for the MSE and

$$\frac{1}{\sqrt{2\pi}}\left(\sigma_1 + \frac{\exp[2H(Y|X)]}{2\pi e\sigma_1}\right) \tag{14}$$

for the MAE. Hence, for a given value of the conditional target entropy, it is possible to obtain feature subsets with different MSE or MAE values. In this setting, mutual information may therefore also fail.

6. Conclusion

The goal of this paper is to study the adequacy of mutual information for feature selection in regression. The conclusion is that mutual information remains optimal in many cases, yet may sometimes also give non-optimal results. On the one hand, mutual information is optimal for commonly assumed estimation error distributions like e.g. the uniform, Laplacian or Gaussian distributions. In such a case, if the estimation error is identically distributed for any $x \in X$, the feature subset with the maximum mutual information is always the feature subset with the lowest MSE or MAE. On the other hand, mutual information may select feature subsets with non-optimal MSE or MAE when e.g. a Student distribution can be assumed for the estimation error. In such a case, feature subsets with identical mutual information values may correspond to different MSE or MAE values, even if the importance of the mutual information failure remains limited in practice.

In practice, it seems that the nature of the estimation errors is an important factor to determine whether mutual information is optimal for feature selection in regression. Hence, any study using mutual information in this context should assess the hypotheses which can be made about the conditional estimation error.

Acknowledgements

The authors thank the ESANN'12 reviewers and attendees and the Neural Networks anonymous reviewers for their fruitful discussions and comments on the subject of this paper, in particular Pierre Dupont, Amaury Lendasse, Fabrice Rossi and Jochen J. Steil.

Gauthier Doquire is funded by a Belgian F.R.I.A. grant.

References

Archambeau, C., Delannay, N., & Verleysen, M. (2008). Mixtures of robust probabilistic principal component analyzers. *Neurocomputing*, *71*(7–9), 1274–1282.

Ash, R. (1990). *Information theory*. Dover Publications.

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

Bellman, R. E. (1961). *Adaptive control processes—a guided tour*. Princeton University Press.

Chen, B., Hu, J., Li, H., & Sun, Z. (2008). Adaptive filtering under maximum mutual information criterion. *Neurocomputing*, *71*(16–18), 3680–3684.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory* (1st ed.). Wiley-Interscience.

Dijck, G. V., & Hulle, M. M. V. (2006). Speeding up the wrapper feature subset selection in regression by mutual information relevance and redundancy analysis. In *Proceedings of the 16th international conference on artificial neural networks* (pp. 31–40). Springer.

Emmert-Streib, F., & Dehmer, M. (2008). *Information theory and statistical learning*. Springer.

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

Frénay, B., Doquire, G., & Verleysen, M. (2012). On the potential inadequacy of mutual information for feature selection. In *Proceedings of ESANN*.

Frénay, B., Doquire, G., & Verleysen, M. (2013). Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification. *Neurocomputing*, *112*, 64–78.

Globerson, A., & Tishby, N. (2003). Sufficient dimensionality reduction. *Journal of Machine Learning Research*, *3*, 1307–1331.

Guo, D., Shamai, S., & Verdú, S. (2005). Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, *51*(4), 1261–1282.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Ihara, S. (1993). *Information theory for continuous system*. World Scientific.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*, 273–324.

Kojadinovic, I., & Wottka, T. (2000). Comparison between a filter and a wrapper approach to variable subset selection in regression problems. In *Proceedings of ESIT*.

Kotz, S., Kozubowski, T., & Podgórski, K. (2001). *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Birkhäuser.

Kozachenko, L. F., & Leonenko, N. (1987). Sample estimate of the entropy of a random vector. *Problems of Information Transmission*, *23*, 95–101.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, *69*, 066138.

Peel, D., & McLachlan, G. J. (2000). Robust mixture modelling using the *t* distribution. *Statistics and Computing*, *10*(4), 339–348.

Psarakis, S., & Panaretos, J. (1990). The folded *t* distribution. *Communications in Statistics: A Theory and Methods*, *19*(7), 2717–2734.

Rossi, F., François, D., Wertz, V., Meurens, M., & Verleysen, M. (2007). Fast selection of spectral variables with *b*-spline compression. *Chemometrics and Intelligent Laboratory Systems*, *86*(2), 208–218. 4.

Rossi, F., Lendasse, A., Francois, D., Wertz, V., & Verleysen, M. (2006). Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *Chemometrics and Intelligent Laboratory Systems*, *80*, 215–226.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

Verleysen, M. (2003). Learning high-dimensional data. In *Limitations and future trends in neural computation, Vol. 186* (pp. 141–162).