A reliable version of choquistic regression based on evidence theory

Sébastien Ramel, Frédéric Pichon, François Delmotte

Version of Record: https://www.sciencedirect.com/science/article/pii/S0950705120304482
Manuscript_424ff623adabc5d059f4c438e9288951

A Reliable Version of Choquistic Regression based on Evidence Theory

Sébastien Ramel, Frédéric Pichon*, François Delmotte

Univ. Artois, UR 3926,
Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A),
F-62400 Béthune, France
{sebastien.ramel,frederic.pichon,francois.delmotte}@univ-artois.fr
*Corresponding author

Abstract

Choquistic regression is an elegant generalisation of logistic regression, which preserves its monotonicity whilst alleviating its linearity. However, much as logistic regression, it lacks self-awareness, that is, an ability to represent the ignorance (*aka* epistemic uncertainty) involved in its predictions, which is crucial in safety-critical classification problems. Recently, an extension of logistic regression was introduced to remedy this issue for this latter classifier. This extension is formalised within evidence theory and relies in particular on a sound method for statistical inference and prediction developed in this framework. In this paper, a similar extension is derived for choquistic regression. The usefulness of the obtained approach is confirmed empirically in classification problems where cautiousness in decision-making is allowed.

Keywords— Belief functions, Choquistic regression, Choquet integral, Logistic regression, Monotonic classification, Reliable classification, Epistemic uncertainty, Nonlinear models.

1 Introduction

In various practical classification problems such as corporate bond rating [46], teaching course evaluation [3] or breast cancer diagnosis [42], it is common to have some prior knowledge that the relation between the input (predictor) variables and the output (target) variable is monotonic, *i.e.*, everything else being equal, the increase of a particular input variable can only increase (or only decrease) the output variable. Accounting for this particular prior knowledge when developing a classification model is important for at least two reasons [47, 26, 17, 1, 4]: (1) it may be beneficial for model induction and in particular to improve the accuracy of predictions; (2) only models satisfying the monotonicity constraint (for short, monotonic models) may be considered acceptable by the end-users. This topic of so-called monotonic classification is not new, yet it has received an increasing interest in the last few years as shown recently in [4].

Focusing on (monotonic) binary classification, where the target variable takes one of two values (the positive class and the negative class), a well-established model for this task is logistic regression [22]. It obtains the probability of the positive class (and thus of the negative class) by modelling the log-odds of the positive class as a linear function of the predictors. Its monotonicity together with its interpretability has contributed to its popularity as underlined in [17]. Yet, logistic regression has at least two limitations, which impedes its use in some real classification problems.

First, the model for the log-odds of the positive class being linear in the predictors, there is a lack of flexibility from a learning point of view with respect to the possibility of interactions between the predictors. This issue can be overcome by replacing the linear model with more complex models that are nonlinear in the predictors [23]. However, this increased flexibility may come at the expense of losing monotonicity and of affecting interpretability. For instance, kernel logistic regression [52] with polynomial and Gaussian kernels are flexible extensions of logistic regression but they are not necessarily monotone [17]. A notable exception, though, is the proposal of Fallah Tehrani *et al.* [17], which uses the Choquet integral [6] for modelling the log-odds of the positive class. The obtained classification model, called choquistic regression, generalises the logistic regression, guarantees monotonicity and provides flexibility in terms of modelling nonlinear relationships between predictors and the log-odds of the positive class. Moreover, it benefits from measures defined for the Choquet integral making it possible to quantify the importance of each predictor as well as the interaction between predictors, and thus it has some level of interpretability. Needless to say, there is a price to pay for the greater flexibility of choquistic regression, which is a higher computational complexity than that of logistic regression. Yet, this complexity may still be acceptable for some real problems.

Secondly, the uncertainty in a given prediction being modelled by a probability measure (characterised by a single number, such as the probability of the positive class), there is a lack of flexibility from a representational point of view with respect to the different possible sources of the uncertainty [43, 49]. In particular, logistic regression fails to provide a quantification of the ignorance involved in its prediction, ignorance (*aka* epistemic or reducible uncertainty) being the part of the uncertainty caused by a lack of knowledge, such as a limited amount of training data [43]. Basically, logistic regression lacks the ability of "knowing what it knows and what not" or, for short, "self-awareness" [43]. Such an ability may be important in critical classification problems such as medical diagnosis, where it may be used to, *e.g.*, postpone the ultimate decision until further data is acquired. This issue of so-called reliable classification [43, 7] can be addressed by replacing the probabilistic representation of uncertainty with richer uncertainty representations. Two noteworthy proposals in this respect, extending logistic regression, are those of Senge *et al.* [43] and of Minary *et al.* [31]. Both are grounded in probability and statistics, but the former is formalised within the framework of fuzzy preference modelling [19] whereas the latter is formalised within evidence theory (*aka* theory of belief functions or Dempster-Shafer theory) [45]. They were applied to two different problems: medical diagnosis and binary SVM classifier stacking, respectively. From a formal point of view, the uncertainty representation in each of these two methods is characterised by two numbers (instead of the single number of the conventional logistic regression) and makes it possible to quantify the ignorance in a given prediction. Obviously, these finer representations of the uncertainty involve higher computational costs than that of logistic regression, but these costs may still be tolerable in some cases.

Choquistic regression is a sound, elegant and useful generalisation of logistic regression, allowing

flexibility in the modelling of interactions between predictors, while maintaining monotonicity and interpretability. However, it suffers from the same limitation as logistic regression with respect to self-awareness. Therefore, it seems interesting to try and extend to choquistic regression the approach of Senge *et al.* [43] or of Minary *et al.* [31]. In this paper, a first study in this direction is reported, where the approach of Minary *et al.* [31] is extended to choquistic regression. We focused on this approach (rather than the one of Senge *et al.* [43]) because it relies on a recent method for statistical inference and prediction proposed in [25], which has gained momentum and is already quite well developed on both the theoretical and practical sides (see, *e.g.*, [29, 37, 24, 30, 12, 5, 15] for recent results around this method, including a few successful applications). Another reason for this focus is the study conducted in the context of active learning in [41], which provides an additional experimental argument that the quantification of ignorance permitted by the approach of [31] can be a useful measure of the lack of knowledge.

The rest of this paper is organised as follow. The logistic regression model and its generalisation based on the Choquet integral due to [17], are recalled in Section 2. Then, in Section 3, necessary notions of evidence theory are provided and the extension of logistic regression based on this theory, introduced in [31], is presented. This latter extension is carried over to the more general choquistic regression model in Section 4. The resulting evidential extension of choquistic regression is compared experimentally in Section 5 to choquistic regression and to the evidential extension of logistic regression. Finally, Section 6 concludes the paper.

2 Choquistic regression

In this section, logistic regression is first recalled. Then, necessary material on the Choquet integral is provided and an extension of logistic regression relying on this integral is presented.

2.1 Logistic regression

As already mentioned, logistic regression (LR) [22] is a model commonly used in binary classification problems. For an object (instance) with unknown label (class or target) $y \in \mathcal{Y} = \{0, 1\}$ and observed feature (predictor) vector $\boldsymbol{x} = (x_1, \ldots, x_m)^\intercal \in \mathcal{X} = \mathbb{R}^m$, and given a training set $\mathcal{D} = \{\boldsymbol{x}^{(i)}, y^{(i)}\}_{i=1}^n$ of n objects supposed to be an *i.i.d.* sample from an underlying (unknown) probability measure P_{XY}, LR yields a probability $P(y = 1|\boldsymbol{x})$ for the positive class (and thus also of the negative class) given the predictors.

LR assumes a monotone relationship of the probability with respect to the predictors, in the sense that every thing else being equal in the predictors, an increase of the value of an individual predictor x_j can either only increase or only decrease the probability.

More specifically, LR is obtained by assuming that the log-odds of $y = 1$ given the predictor vector $\boldsymbol{x}$ is a linear function of the predictors:

$$\log\left(\frac{P(y=1|\boldsymbol{x})}{P(y=0|\boldsymbol{x})}\right) = \sigma_0 + \boldsymbol{\sigma}_m^\intercal \boldsymbol{x}, \tag{1}$$

where $\sigma_0 \in \mathbb{R}$ is a bias (the intercept) and $\boldsymbol{\sigma}_m = (\sigma_1, \ldots, \sigma_m)^\intercal \in \mathbb{R}^m$ is a vector of regression coefficients. According to Eq. (1), a positive (resp. negative) coefficient $\sigma_j > 0$ implies that an increase of x_j will necessarily increase (resp. decrease) the probability of the positive class. Besides, from (1), one easily obtains

$$\pi_l := P(y = 1|\boldsymbol{x}) = h(\sigma_0 + \boldsymbol{\sigma}_m^\intercal \boldsymbol{x}), \tag{2}$$

where

$$h(z) = (1 + \exp(-z))^{-1}, \quad \forall z \in \mathbb{R},$$

is known as the logistic function.

Let $\boldsymbol{\sigma} = (\sigma_0, \boldsymbol{\sigma}_m) \in \Sigma = \mathbb{R}^{m+1}$ denote the parameters of the LR model defined by (2). The actual values $\hat{\boldsymbol{\sigma}}$ of the parameters $\boldsymbol{\sigma}$ to be used in (2) are determined thanks to the training set $\mathcal{D}$. More precisely, they are chosen as the ones maximising the conditional likelihood function:

$$\begin{aligned} L_{\mathcal{D}}(\boldsymbol{\sigma}) &= \prod_{i=1}^{n} P(y^{(i)}|\boldsymbol{x}^{(i)}) \\ &= \prod_{i=1}^{n} (\pi_l^{(i)})^{y^{(i)}} \cdot (1-\pi_l^{(i)})^{1-y^{(i)}}, \quad \forall \boldsymbol{\sigma} \in \Sigma. \end{aligned} \tag{3}$$

In practice, the conditional log-likelihood function

$$\begin{aligned} \ell_{\mathcal{D}}(\boldsymbol{\sigma}) &= \log L_{\mathcal{D}}(\boldsymbol{\sigma}) \\ &= \sum_{i=1}^{n} y^{(i)} \log \pi_l^{(i)} + (1-y^{(i)}) \log(1-\pi_l^{(i)}), \quad \forall \boldsymbol{\sigma} \in \Sigma, \end{aligned}$$

is maximised to find $\hat{\boldsymbol{\sigma}}$ since it has a more convenient form than (3) and its maximum is also attained for $\boldsymbol{\sigma} = \hat{\boldsymbol{\sigma}}$. Formally, we have:

$$\hat{\boldsymbol{\sigma}} = \arg\max_{\boldsymbol{\sigma} \in \Sigma} \ \ell_{\mathcal{D}}(\boldsymbol{\sigma}), \tag{4}$$

and $\hat{\boldsymbol{\sigma}}$ is called a maximum likelihood estimate (MLE) of $\boldsymbol{\sigma}$.

Example 1. *Let $\mathcal{D}$ be the training set composed of $n = 30$ training instances generated as follows:* 15 *instances for the positive class and* 15 *instances for the negative class have been drawn respectively from the bivariate normal distributions $\mathcal{N}(\mu_1, S_1)$ and $\mathcal{N}(\mu_0, S_0)$, with means $\mu_1 = (1, 1)$ and $\mu_0 = (-1, -1)$, and covariance matrices S_1 and S_0 such that*

$$S_0 = \begin{pmatrix} 4 & 3 \\ 3 & 4 \end{pmatrix} \quad \textit{and} \quad S_1 = \begin{pmatrix} 3 & 2.5 \\ 2.5 & 3 \end{pmatrix}.$$

These instances are illustrated in Figure 1. Let us consider an object (also shown in Fig. 1) with observed feature vector $\boldsymbol{x} = (2, -1)$ and unknown label $y \in \mathcal{Y} = \{0, 1\}$. The probability $P(y = 1|\boldsymbol{x})$ of the positive class for this object given its predictors may then be obtained using LR, which essentially amounts to finding a MLE $\hat{\boldsymbol{\sigma}} = (\hat{\sigma_0}, \hat{\sigma_1}, \hat{\sigma_2})$ of $\boldsymbol{\sigma}$ using (4) *and plugging it in Eq.* (2)*: we find*

$$\hat{\boldsymbol{\sigma}} \approx (0.140,\ 1.222,\ 0.848),$$

which induces $P(y = 1|\boldsymbol{x}) \approx 0.85$.

Using the classical (Bayesian) decision strategy of minimising the expected loss [16], the decision reached for the label y of this object is $y = 1$ in the case where the $0/1$ loss function (a wrong decision costs 1 *and a correct decision costs nothing) is used. More generally, Figure 1 shows the decision boundary of LR in the case of $0/1$ loss, which corresponds to those $\boldsymbol{x} \in \mathcal{X}$ such that $P(y = 1|\boldsymbol{x}) = 0.5$.*

2.2 The Choquet integral

The Choquet integral [6] allows the integration of functions with respect to non-additive measures, also known as capacities.

Definition 1 (Capacity)**.** *Let $C = \{c_1, \cdots, c_m\}$ be a finite set. A capacity is a set function $\mu : 2^C \to [0, 1]$ satisfying monotonicity,* i.e.*, $\mu(A) \leq \mu(B)$ for all $A \subseteq B \subseteq C$, and normalisation,* i.e.*, $\mu(\emptyset) = 0$ and $\mu(C) = 1$.*

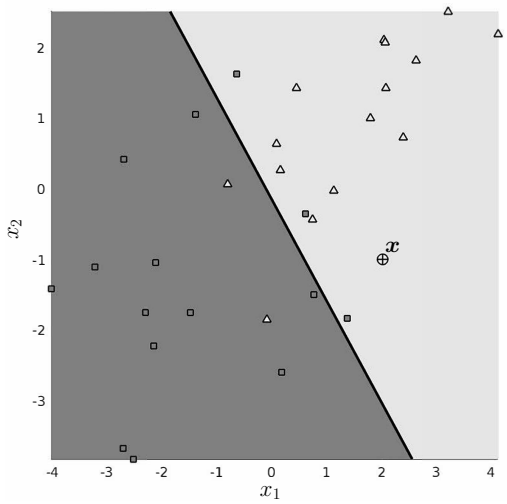


Figure 1: Training set $\mathcal{D}$: instances of the positive class (light grey triangles) and of the negative class (dark grey squares). Object of interest ($\oplus$) with feature vector $\boldsymbol{x} = (2, -1)$. Decision boundary (–) of LR in the case of 0/1 loss, with associated decision regions: $y = 1$ (light grey) and $y = 0$ (dark grey).

Definition 2 (Choquet integral)**.** *Let μ be a capacity on $C = \{c_1, \cdots, c_m\}$. The Choquet integral of a function $g : C \to \mathbb{R}$ with respect to μ is given by*

$$\mathcal{C}_\mu(g) = \sum_{i=1}^{m} \big(g(c_{(i)}) - g(c_{(i-1)})\big) \cdot \mu(A_{(i)}), \tag{5}$$

where $A_{(i)} = \{c_{(i)}, \ldots, c_{(m)}\}$, $g(c_{(0)}) = 0$ and $(\cdot)$ is a permutation of $\{1, \ldots, m\}$ such that $0 \leq g(c_{(1)}) \leq g(c_{(2)}) \leq \ldots \leq g(c_{(m)})$.

Besides its use in classification (recalled in Section 2.3), the Choquet integral plays a role in other domains. For instance, computing the lower expectation of a function of an uncertain variable, when uncertainty is modelled by a belief function (which is a particular kind of non-additive measures for the representation of uncertainty) corresponds to computing the Choquet integral of this function with respect to the belief function [20]. However, above all, its prominent field of application has been that of multiple criteria decision making (MCDM), where it is used as an aggregation operator (the non-additivity allowing to account for interactions between criteria) [21] as briefly recalled in the following.

When using the Choquet integral in MCDM, C represents a set of relevant criteria to compare different alternatives, $\mu(A)$ is interpreted as the importance of the subset of criteria $A \subseteq C$, $g(c_i) \in [0, 1]$ is the utility of a given alternative with respect to criterion c_i indicating the degree to which c_i is satisfied for this alternative, and $\mathcal{C}_\mu(g)$ represents the overall evaluation (satisfaction) of this alternative. For instance, if one is contemplating buying a house, the alternatives are the different houses on offer, the set of criteria might be $C = \{$`living area`, `garden size`, `number of rooms`$\}$ and $\mu(A)$ is the evaluation of an alternative satisfying criteria A (and not satisfying $C \backslash A$). Besides, interactions between criteria are accounted for through μ: if subsets of criteria $A \subseteq C$ and $B \subseteq C$ can be seen as complementary (*e.g.*, $A = \{$`living area`$\}$ and $B = \{$`garden size`$\}$ since both areas are important on the final decision) then this is formally expressed by $\mu(A \cup B) > \mu(A) + \mu(B)$ (positive interaction), whereas if A and B are redundant (*e.g.*, $A = \{$`living area`$\}$ and $B = \{$`number of rooms`$\}$ since either of the two may suffice), this is represented by $\mu(A \cup B) < \mu(A) + \mu(B)$ (negative interaction).

The Choquet integral admits a useful representation based on the so-called Möbius transform.

Definition 3 (Möbius transform)**.** *The Möbius transform m_μ of a capacity μ is given by*

$$m_\mu(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|} \mu(B), \forall A \subseteq C.$$

Remarking that $m_\mu(\emptyset) = 0$, we denote by $\boldsymbol{m}_\mu$ the vector of size 2^{m-1} whose elements are the values $m_\mu(A) \in \mathbb{R}$, $\emptyset \neq A \subseteq C$, ordered according to some arbitrary order on the subsets of C.

A capacity μ can be recovered from its Möbius transform m_μ by:

$$\mu(A) = \sum_{B \subseteq A} m_\mu(B), \forall A \subseteq C. \tag{6}$$

Using (6), it is easy to show that the Choquet integral (5) of a function g with respect to the capacity μ satisfies:

$$\mathcal{C}_\mu(g) = \sum_{A \subseteq C} m_\mu(A) \cdot \min_{c_i \in A} g(c_i). \tag{7}$$

An interesting property that a capacity μ may satisfy is k-order additivity (or, simply, k-additivity): if k is the smallest integer such that $m_\mu(A) = 0$ for all $A \subseteq C$ with $|A| > k$, then μ is said to be k-additive. In MCDM, this property means an absence of interaction between subsets of criteria $A, B \subset C$ such that $|A| > k$ and $|B| > k$. It also implies that μ is then characterised by less than the 2^{m-1} values required in the general case.

2.3 The choquistic regression model

In [17], it is proposed to extend the LR model by replacing $\sigma_0 + \boldsymbol{\sigma}_m^\intercal \boldsymbol{x}$ in (1) by

$$\gamma \left(\mathcal{C}_\mu(g_{\boldsymbol{x}}) - \beta \right), \tag{8}$$

which yields

$$\pi_c := P(y = 1|\boldsymbol{x}) = h(\gamma \left(\mathcal{C}_\mu(g_{\boldsymbol{x}}) - \beta \right)), \tag{9}$$

where γ and β are two parameters such that $\gamma > 0$ and $\beta \in [0, 1]$, and where $\mathcal{C}_\mu(g_{\boldsymbol{x}})$ is the Choquet integral with respect to measure μ of the function

$$g_{\boldsymbol{x}} : C \to [0, 1]$$

that maps, for a given object with feature value vector $\boldsymbol{x} = (x_1, \ldots, x_m)^\intercal$, each feature c_i (viewing C as the set of features describing objects) to a value $\tilde{x}_i = g_{\boldsymbol{x}}(c_i) \in [0, 1]$ corresponding to a normalisation of the feature value x_i.

Assuming that some prior knowledge on the monotonicity of the classification problem at hand is available and, more specifically, that the direction of the influence of each input feature x_i on the probability of the positive class is known, then the following normalisation can be used [17]: if the influence is positive (increasing), then

$$\tilde{x}_i = \frac{x_i - m_i}{M_i - m_i}, \tag{10}$$

with m_i and M_i the lower and upper bounds for x_i, which are either known or estimated from the learning set $\mathcal{D}$ as $m_i = \min_{1 \leq j \leq n} x_i^{(j)}$ and $M_i = \max_{1 \leq j \leq n} x_i^{(j)}$; and if the influence is negative (decreasing), then $\tilde{x}_i = (M_i - x_i)/(M_i - m_i)$. In the case where the direction of the influence of predictor x_i is actually not known, then it is estimated from the data using LR [17]: if $\hat{\sigma}_i > 0$ then the influence is assumed to be positive, otherwise it is considered to be negative.

In a nutshell, this approach, called choquistic regression (CR), proceeds in two steps. First, an aggregated value $\mathcal{C}_\mu(g_{\boldsymbol{x}}) \in [0, 1]$ is obtained for an instance $\boldsymbol{x}$. Then, this value is compared to the threshold β. If $\mathcal{C}_\mu(g_{\boldsymbol{x}}) > \beta$, then $P(y = 1|\boldsymbol{x}) > 0.5$ ("the decision tends to be positive" [17]), whereas if $\mathcal{C}_\mu(g_{\boldsymbol{x}}) < \beta$, then $P(y = 1|\boldsymbol{x}) < 0.5$ (the decision "tends to be negative" [17]). Furthermore, the

parameter γ acts as a scaling factor. If $\mathcal{C}_\mu(g_{\boldsymbol{x}}) > \beta$, then as γ increases, $P(y = 1|\boldsymbol{x})$ increases, whereas if $\mathcal{C}_\mu(g_{\boldsymbol{x}}) < \beta$, then as γ increases, $P(y = 1|\boldsymbol{x})$ decreases. In addition, let us remark that this model is a proper generalisation of LR (we refer the interested reader to [17, Section 5.3]).

Let $\boldsymbol{\upsilon} = (\gamma, \boldsymbol{m}_\mu, \beta)$ denote the parameters of the CR model defined by (9), where m_μ is the Möbius transform of μ. Similarly as in the case of LR, it is proposed in [17] to determine the actual values $\hat{\boldsymbol{\upsilon}}$ for the parameters $\boldsymbol{\upsilon}$ using the maximum likelihood principle. The conditional likelihood of the CR parameters is given by

$$L_{\mathcal{D}}(\boldsymbol{\upsilon}) = \prod_{i=1}^{n} (\pi_c^{(i)})^{y^{(i)}} \cdot (1 - \pi_c^{(i)})^{1-y^{(i)}}. \tag{11}$$

Hence, the conditional log-likelihood function can be written as:

$$\ell_{\mathcal{D}}(\boldsymbol{\upsilon}) = \sum_{i=1}^{n} y^{(i)} \log \pi_c^{(i)} + (1 - y^{(i)}) \log(1 - \pi_c^{(i)}). \tag{12}$$

Unlike the coefficients $\boldsymbol{\sigma}$ of LR, the parameters $\boldsymbol{\upsilon}$ of CR must respect some constraints and in particular m_μ must be the Möbius transform of some proper capacity μ. Therefore the maximisation of the conditional log-likelihood (12) is actually a constrained optimisation problem, which may be formally written as [17]:

$$\hat{\boldsymbol{\upsilon}} = \arg\max_{\boldsymbol{\upsilon} \in \Upsilon} \ \ell_{\mathcal{D}}(\boldsymbol{\upsilon}) \tag{13}$$

with Υ the set composed of the vectors $(\gamma, \boldsymbol{m}_\mu, \beta)$ satisfying

$$\beta \in [0, 1], \quad \gamma \in \mathbb{R}_{>0}, \quad \sum_{A \subseteq C} m_\mu(A) = 1 \tag{14}$$

and

$$\sum_{B \subseteq A \setminus \{c_i\}} m_\mu(B \cup \{c_i\}) \geq 0, \quad \forall A \subseteq C, \ \forall c_i \in C. \tag{15}$$

Example 2. *Continuing Example 1, the probability $P(y = 1|\boldsymbol{x})$ of the positive class for the object with feature vector $\boldsymbol{x} = (2, -1)$ may also be obtained using CR, which amounts to finding a MLE $\hat{\boldsymbol{\upsilon}}$ of $\boldsymbol{\upsilon}$ using* (13) *and plugging it in* (9)*: we find*

$$\widehat{m_\mu}(\{1\}) \approx 0.020, \quad \widehat{m_\mu}(\{2\}) = 0, \quad \widehat{m_\mu}(\{1, 2\}) \approx 0.980, \quad \hat{\gamma} \approx 15.077, \quad \hat{\beta} \approx 0.426,$$

which induces $P(y = 1|\boldsymbol{x}) \approx 0.59$. Similarly as for LR, the decision reached for the label y of this object using CR is $y = 1$ in the case of $0/1$ loss. More generally, Figure 2 shows the decision boundary of CR in the case of $0/1$ loss.

Let us remark that CR is quite a flexible model and runs thus the risk of overfitting the data. A solution proposed in [17] to mitigate this issue is to restrict the capacity μ to be k-additive, for some $k < m$ determined through cross-validation.

3 Evidential logistic regression

An extension of LR based on evidence theory is recalled in this section. For the paper to be self-contained, necessary notions – in particular an approach for statistical inference and prediction – of the theory of evidence on which this extension relies, are provided first.

3.1 Basic notions of the theory of evidence

Evidence theory is a framework for uncertainty modelling and reasoning. Let Y be a variable whose actual value y belongs to some finite set $\mathcal{Y} = \{y_1, \ldots, y_K\}$ (called the frame of discernment). Uncertainty about y is represented in this theory by a mapping $m^{\mathcal{Y}} : 2^{\mathcal{Y}} \to [0, 1]$, called mass function, such that

$$\sum_{A \subseteq \mathcal{Y}} m^{\mathcal{Y}}(A) = 1, \tag{16}$$

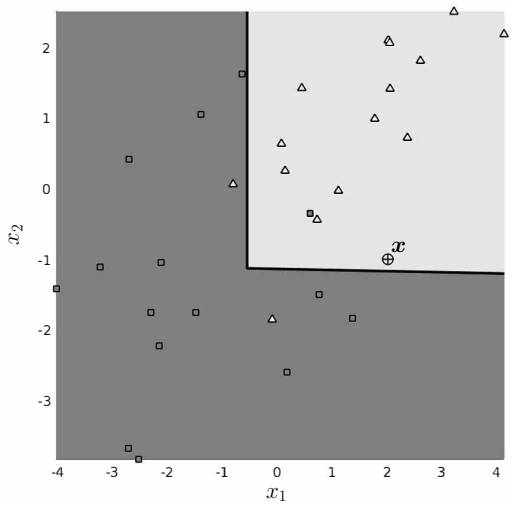


Figure 2: Decision boundary (−) of CR in the case of 0/1 loss.

and $m^{\mathcal{Y}}(\emptyset) = 0$. The quantity $m^{\mathcal{Y}}(A)$ is interpreted as the belief allocated exactly to the hypothesis $y \in A$ and nothing more specific. In addition, any $A \subseteq \mathcal{Y}$ such that $m^{\mathcal{Y}}(A) > 0$ is called a focal set of $m^{\mathcal{Y}}$.

Several functions are in one-to-one correspondence with $m^{\mathcal{Y}}$ and, in particular, the belief and plausibility functions defined, respectively, for all $A \subseteq \mathcal{Y}$ as

$$Bel^{\mathcal{Y}}(A) = \sum_{B \subseteq A} m^{\mathcal{Y}}(B), \quad \text{and} \quad Pl^{\mathcal{Y}}(A) = \sum_{B \cap A \neq \emptyset} m^{\mathcal{Y}}(B). \tag{17}$$

The degree of belief $Bel^{\mathcal{Y}}(A)$ is interpreted as the amount of evidence strictly supporting $y \in A$, while the plausibility $Pl^{\mathcal{Y}}(A)$ represents the amount of evidence not contradicting $y \in A$. Due to the one-to-one correspondence between these functions, $m^{\mathcal{Y}}$ can be recovered from, *e.g.*, $Bel^{\mathcal{Y}}$. We have

$$m^{\mathcal{Y}}(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|} Bel^{\mathcal{Y}}(B), \quad \forall A \subseteq \mathcal{Y}. \tag{18}$$

The contour function $pl^{\mathcal{Y}}$ is defined as the plausibility function restricted to the singletons, *i.e.*, $pl^{\mathcal{Y}}(y_i) = Pl^{\mathcal{Y}}(\{y_i\})$, for all $y_i \in \mathcal{Y}$. When its focal sets are nested, a mass function $m^{\mathcal{Y}}$ is said to be consonant and it is fully characterised by its contour function. Its associated plausibility function can then be recovered as follows:

$$Pl^{\mathcal{Y}}(A) = \sup_{y_i \in A} pl^{\mathcal{Y}}(y_i), \quad \forall A \subseteq \mathcal{Y}. \tag{19}$$

There exist several strategies to make a decision about the actual value y of Y given a mass function $m^{\mathcal{Y}}$ on $\mathcal{Y}$ representing uncertainty about y [10, 13, 28]. Some yield systematically a precise decision, *i.e.*, $y = y_i$ for some $y_i \in \mathcal{Y}$, while other may yield only an imprecise decision, *i.e.*, $y \in A$ for some $A \subseteq \mathcal{Y}$, when there is too much uncertainty. In the particular case where the frame of discernment is binary (by convention, in this paper, $\mathcal{Y} = \{0, 1\}$) and 0/1 loss is used, then the classical precise decision strategies (*i.e.*, minimising the pignistic, or lower, or upper expected loss, see [10]) amount to the same decision rule:

$$y = \begin{cases} 1 \text{ if } m^{\mathcal{Y}}(\{1\}) > m^{\mathcal{Y}}(\{0\}), \\ 0 \text{ otherwise.} \end{cases} \tag{20}$$

Among the imprecise decision strategies, the interval dominance decision rule has recently shown its interest in exploiting evidential classifiers [14]. In the binary case and with 0/1 loss, this rule reads as

follows:

$$y = \begin{cases} 1 \text{ if } Bel^{\mathcal{Y}}(\{1\}) > Pl^{\mathcal{Y}}(\{0\}), \\ 0 \text{ if } Bel^{\mathcal{Y}}(\{0\}) > Pl^{\mathcal{Y}}(\{1\}), \end{cases} \tag{21}$$

and $y \in \{0, 1\}$ otherwise.

3.2 Belief function-based approach to statistical inference and prediction

Shafer [45] introduced a belief function-based approach to statistical inference, which was subsequently justified by Denoeux [11]. Given an observed realisation $z \in \mathcal{Z}$ of some random quantity $Z \sim P_Z(\cdot;\theta)$ with $\theta \in \Theta$ the unknown parameter of interest, this approach represents knowledge about θ by a consonant belief function Bel_z^Θ whose contour function pl_z^Θ is the relative likelihood:

$$pl_z^\Theta(\theta) = \frac{L_z(\theta)}{L_z(\hat{\theta})}, \quad \forall\, \theta \in \Theta,$$

where $L_z(\theta) = P_Z(z;\theta)$, $\hat{\theta}$ is a MLE of θ and it is assumed that $L_z(\hat{\theta}) < \infty$.

Bel_z^Θ is called the likelihood-based belief function. Its focal sets are the level-sets of pl_z^Θ defined as

$$\Gamma_z(u) = \{\theta \in \Theta \mid pl_z^\Theta(\theta) \geq u\}, \tag{22}$$

for $u \in [0, 1]$. Moreover, Bel_z^Θ can be regarded as being induced by the random set [34] $\Gamma_z(U)$ with $U \sim \mathcal{U}([0, 1])$, in the sense that

$$Bel_z^\Theta(A) = P_U(\{u \in [0, 1] | \Gamma_z(u) \subseteq A\}), \quad \forall A \subseteq \Theta.$$

In [25, 24], this likelihood-based approach to statistical inference was extended to the prediction problem, which consists in making statements about a not-yet-observed realisation $y \in \mathcal{Y}$ of some random quantity $Y \sim P_Y(\cdot;\theta)$ given knowledge about θ obtained by observing z (represented here by Bel_z^Θ). The extension relies on Dempster's sampling model [8], which expresses Y as a function φ of the parameter θ and some unobserved variable V with known probability distribution P_V independent of θ:

$$Y = \varphi(\theta, V). \tag{23}$$

In [25, 24] (see, also, [12]), the function φ is obtained by inverting the cdf of Y and $V \sim \mathcal{U}([0, 1])$.

In this scheme, variables U and V are independent and thus the distribution $P_{U,V}$ of (U, V) is the uniform distribution on $[0, 1]^2$. Besides, for any $(u, v) \in [0, 1]^2$, we can assert from (22) and (23) that $Y \in \varphi(\Gamma_z(u), v)$. Accordingly, knowledge about the future realisation y given the observed data z may be represented by the belief function $Bel_z^\mathcal{Y}$ induced by the random set $\varphi(\Gamma_z(U), V)$ and defined as [25, 24]

$$Bel_z^\mathcal{Y}(A) = P_{U,V}\left(\{(u, v) \in [0, 1]^2 \mid \varphi(\Gamma_z(u), v) \subseteq A\}\right), \quad \forall A \subseteq \mathcal{Y}. \tag{24}$$

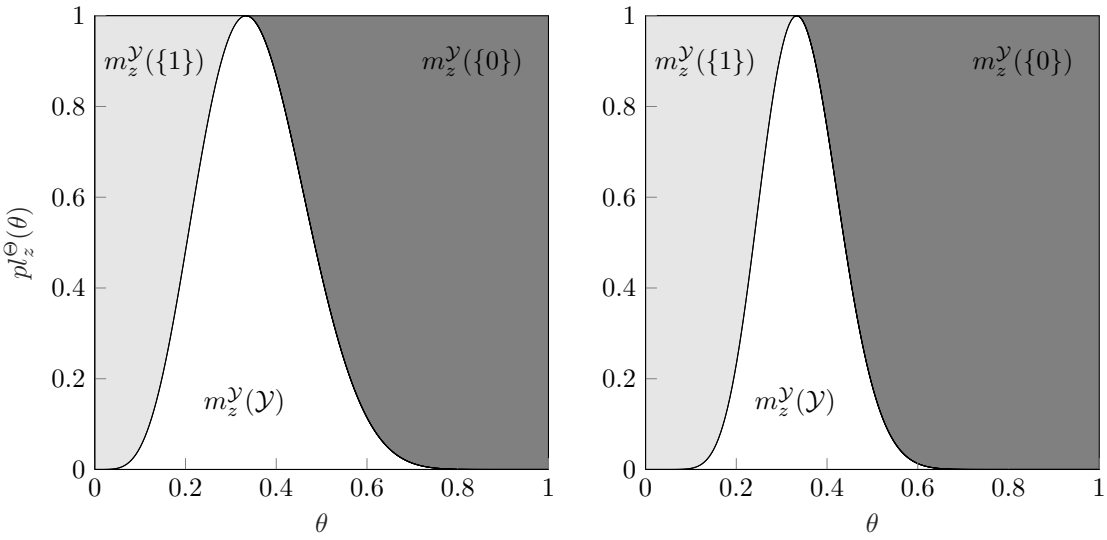
$Bel_z^\mathcal{Y}$ is called the predictive belief function. Its associated plausibility function $Pl_z^\mathcal{Y}$ is defined by

$$Pl_z^\mathcal{Y}(A) = P_{U,V}\left(\{(u, v) \in [0, 1]^2 \mid \varphi(\Gamma_z(u), v) \cap A \neq \emptyset\}\right), \quad \forall A \subseteq \mathcal{Y}. \tag{25}$$

This approach to statistical inference and prediction is illustrated by Example 3, where Y is a Bernoulli variable, as this is the case on which the evidential extension of LR recalled in the next section relies.

Example 3. *Let $Z \sim \mathcal{B}(n, \theta)$,* i.e.*, Z is a random variable following a binomial distribution with parameters $n \in \mathbb{N}$ and $\theta \in [0, 1]$. Assume we have observed $z \in \mathbb{N}$ successes out of the n trials. Knowledge about θ given z is then represented by the likelihood-based belief function Bel_z^Θ with contour function defined by:*

$$pl_z^\Theta(\theta) = \frac{\theta^z(1-\theta)^{n-z}}{\hat{\theta}^z(1-\hat{\theta})^{n-z}} = \left(\frac{\theta}{\hat{\theta}}\right)^{n\hat{\theta}} \cdot \left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n(1-\hat{\theta})}, \quad \forall \theta \in [0, 1],$$



(a) $n = 15$ (b) $n = 30$

Figure 3: Contour functions for the binomial distribution with $\hat{\theta} = \frac{1}{3}$ and $n \in \{15, 30\}$, and induced predictive mass functions for the next trial.

where $\hat{\theta} = z/n$ is the MLE of θ. Figures 3a and 3b show pl_z^{Θ} for $\hat{\theta} = \frac{1}{3}$ and for, respectively, $n = 15$ and $n = 30$. As can be seen, the area under the contour function pl_z^{Θ} reduces as n increases.

Let $Y \sim \mathcal{B}(\theta)$, i.e., *$Y$ is a binary ($\mathcal{Y} = \{0, 1\}$) random variable following a Bernoulli distribution with parameter $\theta \in [0, 1]$. Variable Y can be expressed in the form* (23) *as*

$$Y = \varphi(\theta, V) = \begin{cases} 1, & \text{if } V \leq \theta, \\ 0, & \text{otherwise,} \end{cases}$$

with $V \sim \mathcal{U}([0, 1])$.

Using the fact that pl_z^{Θ} is continuous and unimodal (as illustrated by Fig. 3), it can be shown (see [25, Example 2]) that the expressions of the predictive belief and predictive plausibility that $y = 1$, given by (24) *and* (25) *for $A = \{1\}$, reduce respectively to*

$$Bel_z^{\mathcal{Y}}(\{1\}) = \hat{\theta} - \int_0^{\hat{\theta}} pl_z^{\Theta}(\theta) d\theta, \tag{26}$$

$$Pl_z^{\mathcal{Y}}(\{1\}) = \hat{\theta} + \int_{\hat{\theta}}^{1} pl_z^{\Theta}(\theta) d\theta. \tag{27}$$

For instance, for $n = 15$ and $\hat{\theta} = \frac{1}{3}$, we find $Bel_z^{\mathcal{Y}}(\{1\}) \approx 0.20$ and $Pl_z^{\mathcal{Y}}(\{1\}) \approx 0.49$.

Equivalently, in terms of the predictive mass function $m_z^{\mathcal{Y}}$, we have using Eqs. (16)-(18)

$$\begin{aligned} m_z^{\mathcal{Y}}(\{1\}) &= Bel_z^{\mathcal{Y}}(\{1\}), \\ m_z^{\mathcal{Y}}(\{0\}) &= 1 - Pl_z^{\mathcal{Y}}(\{1\}), \\ m_z^{\mathcal{Y}}(\mathcal{Y}) &= Pl_z^{\mathcal{Y}}(\{1\}) - Bel_z^{\mathcal{Y}}(\{1\}) \\ &= \int_0^1 pl_z^{\Theta}(\theta) d\theta. \end{aligned}$$

Each mass $m_z^{\mathcal{Y}}(A)$, $A \subseteq \mathcal{Y}$, corresponds to a particular area with respect to the function pl_z^{Θ}, as illustrated by Fig. 3. In particular, $m_z^{\mathcal{Y}}(\mathcal{Y})$, which represents the amount of belief, called ignorance [25], that cannot

be committed to any specific hypothesis, is equal to the area under the contour function pl_z^{Θ}. *Let us remark that the size of this area tends to* 0 *as* n *tends to infinity [25]. In other words,* $m_z^{\mathcal{Y}}(\mathcal{Y})$ *reflects the amount of data: the more data there are, the less ignorance there is.*

3.3 The evidential logistic regression model

Following previous work from Xu *et al.* [49], Minary *et al.* [31] introduced a belief function-based extension of LR, called hereafter evidential logistic regression (ELR), which they used as a meta-classifier to combine classifiers' outputs. In ELR, the label $y \in \mathcal{Y} = \{0, 1\}$ of an object whose feature vector $\boldsymbol{x}$ has been observed, is seen as the realisation of a binary random variable Y following a Bernoulli distribution with parameter $\theta = h(\sigma_0 + \boldsymbol{\sigma}_m^{\intercal}\boldsymbol{x})$, where there is some uncertainty on $\boldsymbol{\sigma} = (\sigma_0, \boldsymbol{\sigma}_m)$ induced by the observation of the training data $\mathcal{D}$. The representation of the uncertainty on $\boldsymbol{\sigma}$ and the subsequent prediction of Y are carried out using the approach for statistical inference and prediction recalled in the preceding section, therefore yielding a belief degree $Bel_{\mathcal{D},\boldsymbol{x}}^{\mathcal{Y}}(\{1\})$ and a plausibility degree $Pl_{\mathcal{D},\boldsymbol{x}}^{\mathcal{Y}}(\{1\})$ (rather than a probability $P(y = 1|\boldsymbol{x})$ as in LR) for the positive class given the predictors.

More precisely, Minary *et al.* [31] represent uncertainty on $\boldsymbol{\sigma}$ by the likelihood-based (consonant) belief function $Bel_{\mathcal{D}}^{\Sigma}$ whose contour function is defined by

$$pl_{\mathcal{D}}^{\Sigma}(\boldsymbol{\sigma}) = \frac{L_{\mathcal{D}}(\boldsymbol{\sigma})}{L_{\mathcal{D}}(\hat{\boldsymbol{\sigma}})}, \quad \forall \boldsymbol{\sigma} \in \Sigma, \tag{28}$$

and whose corresponding plausibility function is obtained as

$$Pl_{\mathcal{D}}^{\Sigma}(A) = \sup_{\boldsymbol{\sigma} \in A} pl_{\mathcal{D}}^{\Sigma}(\boldsymbol{\sigma}), \quad \forall A \subseteq \Sigma.$$

Then, the uncertainty with respect to θ induced by the uncertainty on $\boldsymbol{\sigma}$ and the observed feature vector $\boldsymbol{x}$, is represented by a consonant belief function $Bel_{\mathcal{D},\boldsymbol{x}}^{\Theta}$ with contour function $pl_{\mathcal{D},\boldsymbol{x}}^{\Theta}$ obtained as

$$pl_{\mathcal{D},\boldsymbol{x}}^{\Theta}(\theta) = \begin{cases} 0 & \text{if } \theta \in \{0, 1\}, \\ Pl_{\mathcal{D}}^{\Sigma}\left(\{\boldsymbol{\sigma} \in \Sigma \mid \theta = h(\sigma_0 + \boldsymbol{\sigma}_m^{\intercal}\boldsymbol{x})\}\right) & \text{otherwise,} \end{cases}$$

where

$$\begin{aligned} \{\boldsymbol{\sigma} \in \Sigma \mid \theta = h(\sigma_0 + \boldsymbol{\sigma}_m^{\intercal}\boldsymbol{x})\} &= \left\{\boldsymbol{\sigma} \in \Sigma \mid -(\sigma_0 + \boldsymbol{\sigma}_m^{\intercal}\boldsymbol{x}) = \ln(\theta^{-1} - 1)\right\} \\ &= \left\{\boldsymbol{\sigma} \in \Sigma \mid \sigma_0 = -\ln(\theta^{-1} - 1) - \boldsymbol{\sigma}_m^{\intercal}\boldsymbol{x}\right\}, \end{aligned}$$

which yields

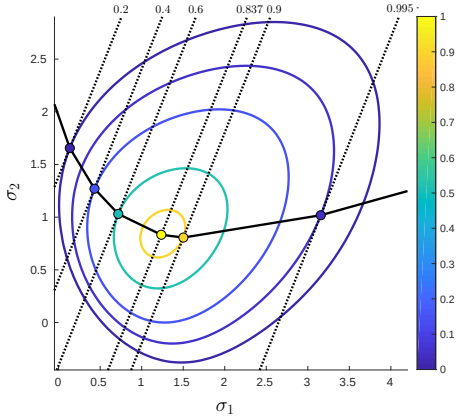
$$pl_{\mathcal{D},\boldsymbol{x}}^{\Theta}(\theta) = \sup_{\boldsymbol{\sigma}_m \in \mathbb{R}^m} pl_{\mathcal{D}}^{\Sigma}\left(-\ln(\theta^{-1} - 1) - \boldsymbol{\sigma}_m^{\intercal}\boldsymbol{x}, \boldsymbol{\sigma}_m\right), \quad \forall \theta \in (0, 1). \tag{29}$$

The value $pl_{\mathcal{D},\boldsymbol{x}}^{\Theta}(\theta)$, $\theta \in (0, 1)$, can be obtained by an iterative maximisation algorithm.

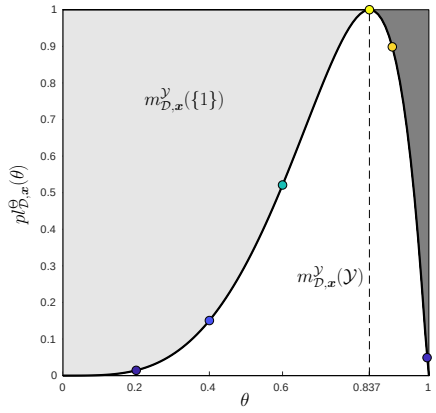
Finally, the predictive belief $Bel_{\mathcal{D},\boldsymbol{x}}^{\mathcal{Y}}(\{1\})$ and predictive plausibility $Pl_{\mathcal{D},\boldsymbol{x}}^{\mathcal{Y}}(\{1\})$ of the positive class given the predictors are obtained from the contour function $pl_{\mathcal{D},\boldsymbol{x}}^{\Theta}$ using Eqs. (26) and (27), respectively. Note that, using Eqs. (16)-(18), the uncertainty with respect to the label y of an object whose feature vector $\boldsymbol{x}$ has been observed, can then be equivalently characterized by some other pairs of quantities than the belief degree $Bel_{\mathcal{D},\boldsymbol{x}}^{\mathcal{Y}}(\{1\})$ and the plausibility degree $Pl_{\mathcal{D},\boldsymbol{x}}^{\mathcal{Y}}(\{1\})$, such as the mass $m_{\mathcal{D},\boldsymbol{x}}^{\mathcal{Y}}(\{1\})$ supporting the positive class together with the mass $m_{\mathcal{D},\boldsymbol{x}}^{\mathcal{Y}}(\mathcal{Y})$ representing the amount of ignorance.

Example 4. *In order to be able to provide graphical representations of the inner workings of ELR, we will consider, only in this example, a slightly simplified version of ELR where a degree of freedom of the model is removed. Specifically, we consider the case where the LR extended by ELR is LR with the bias term* σ_0 *fixed (arbitrarily) to* $\sigma_0 = 0$. *In this case, considering the classification problem with training dataset* $\mathcal{D}$ *of Example 1, LR is defined with coefficients* $\boldsymbol{\sigma}_m = (\sigma_1, \sigma_2)^{\intercal} \in \mathbb{R}^2$.

The steps to obtain the predictive belief $Bel_{\mathcal{D},\boldsymbol{x}}^{\mathcal{Y}}(\{1\})$ *and predictive plausibility* $Pl_{\mathcal{D},\boldsymbol{x}}^{\mathcal{Y}}(\{1\})$ *for the object with observed feature vector* $\boldsymbol{x} = (2, -1)$ *are now illustrated, where the restriction* $\sigma_0 = 0$ *holds throughout the rest of the example.*



(a) Level sets of the contour function $pl^{\Sigma}_{\mathcal{D}}$. Lines $\sigma_2 = (-\ln(\theta^{-1}-1) - \sigma_1 x_1)/x_2$ for $\boldsymbol{x} = (x_1, x_2) = (2, -1)$ and $\theta \in \{0.2, 0.4, 0.6, 0.837, 0.9, 0.995\}$, with the maximum on each line (···○···) .



(b) Contour function $pl^{\Theta}_{\mathcal{D},\boldsymbol{x}}(\theta)$ for $\boldsymbol{x} = (2, -1)$. Associated predictive mass function $m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}$. Maximums (○) found for different values of θ in Fig. 4a.

Figure 4: Computation of the contour function on Θ involved in ELR.

The levels sets of the contour function $pl^{\Sigma}_{\mathcal{D}}$ defined by (28) *are shown in Figure 4a. From* (29) *(and $\sigma_0 = 0$), the value of $pl^{\Theta}_{\mathcal{D},\boldsymbol{x}}(\theta)$ is defined as the maximum of $pl^{\Sigma}_{\mathcal{D}}$ along the line :*

$$\sigma_2 = (-\ln(\theta^{-1} - 1) - \sigma_1 x_1)/x_2.$$

This line is shown for different values of θ in Figure 4a. Figure 4b shows the resulting contour function $pl^{\Theta}_{\mathcal{D},\boldsymbol{x}}$ for feature vector $\boldsymbol{x} = (2, -1)$, from which $Bel^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{1\})$ and $Pl^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{1\})$ are obtained using Eqs. (26) *and* (27) *respectively. We find:*

$$Bel^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{1\}) = 0.5703, \quad Pl^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{1\}) = 0.9517.$$

Equivalently, we have

$$m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{1\}) = 0.5703, \quad m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\mathcal{Y}) = 0.3815$$

(and $m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{0\}) = 0.0483$ using (16)*).*

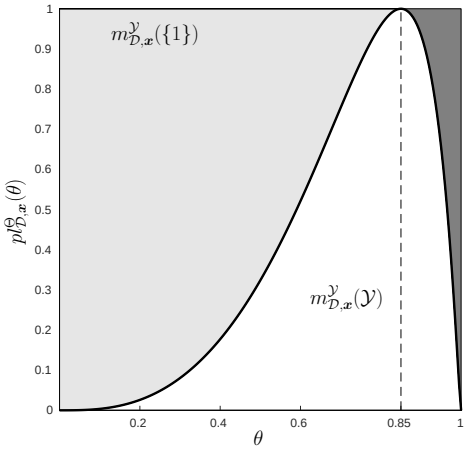
Example 5. *Relaxing the restriction on the bias parameter (*i.e.*, the restriction $\sigma_0 = 0$) used in Example 4 solely for the sake of illustration, let us provide the actual results obtained when one applies ELR to the data from Example 1. Figure 5a shows the contour function $pl^{\Theta}_{\mathcal{D},\boldsymbol{x}}$ for the object with observed feature vector $\boldsymbol{x} = (2, -1)$ from which $Bel^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{1\})$ and $Pl^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{1\})$ are obtained using Eqs.* (26) *and* (27) *respectively. We find:*

$$Bel^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{1\}) = 0.5655, \quad Pl^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{1\}) = 0.9578,$$

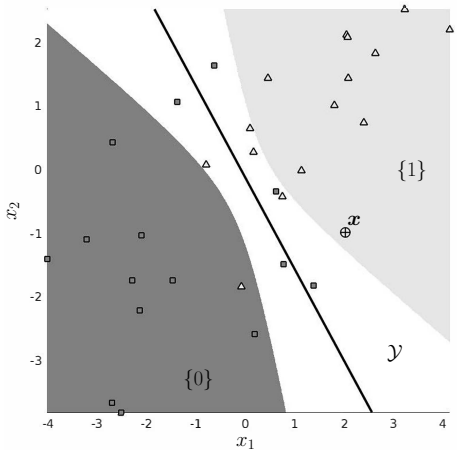
or, equivalently,

$$m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{1\}) = 0.5655, \quad m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\mathcal{Y}) = 0.3923, \quad m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{0\}) = 0.0422.$$

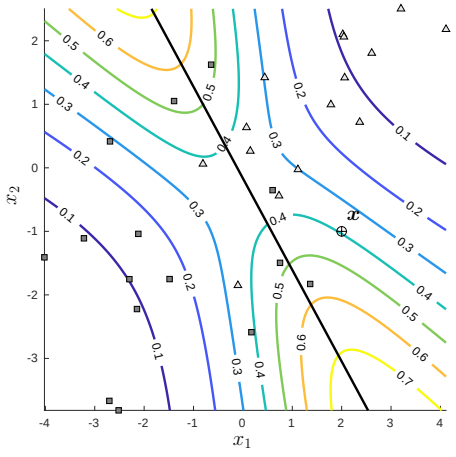
Using the interval dominance decision rule with $m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}$ and 0/1 loss, one obtains that the decision for the label y of the object having feature vector $\boldsymbol{x} = (2, -1)$ is $y = 1$. For this instance, the uncertainty is thus too low to yield an imprecise decision. A precise decision is reached, which is the same decision as the one obtained with LR. Nonetheless, let us remark that in general, decisions made using LR and ELR may differ as illustrated in Figure 5b, which shows the decision regions (in the case of 0/1 loss) of ELR and in particular the fact that it may yield imprecise decisions. For completeness, Figure 5c shows the degree of ignorance yielded by ELR for any $\boldsymbol{x} \in \mathcal{X}$; we can remark that the areas where ignorance is the greatest correspond to the areas where imprecise decisions are made.



(a) Contour function $pl^{\Theta}_{\mathcal{D},\boldsymbol{x}}(\theta)$ for $\boldsymbol{x} = (2, -1)$. Associated predictive mass function $m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}$.



(b) Training instances $\mathcal{D}$. Object of interest ($\oplus$). Decision boundary (–) of LR. Decision regions of ELR according to the interval dominance rule: $y = 1$ (light grey), $y = 0$ (dark grey) and $y \in \mathcal{Y} = \{0, 1\}$ (white).



(c) Training instances $\mathcal{D}$. Decision boundary (–) of LR. Contour lines of the ignorance $m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\mathcal{Y})$ for any $\boldsymbol{x} \in \mathcal{X}$.

Figure 5: Evidential logistic regression.

4 Evidential choquistic regression

In this section, a belief function-based extension of the CR model is proposed, following the approach of Minary *et al.* [31] to extend the LR model that was recalled in Section 3.3. This extension of CR, derived in Section 4.2, is presented using a reparameterisation of CR in a linear form provided in Section 4.1. As will be seen, this reparameterisation offers several advantages. Furthermore, in Section 4.3, the computational complexity of this extension of CR is analysed.

4.1 Reparameterised choquistic regression

The CR model may be rewritten in a linear form as follows. From (7), expression (8) verifies

$$\begin{aligned}\gamma\left(\mathcal{C}_\mu(g_{\boldsymbol{x}})-\beta\right) &= \gamma\left(\sum_{A\subseteq C} m_\mu(A)\cdot\min_{c_i\in A} g_{\boldsymbol{x}}(c_i)-\beta\right)\\ &= -\gamma\beta+\sum_{A\subseteq C}\gamma\cdot m_\mu(A)\cdot\min_{c_i\in A} g_{\boldsymbol{x}}(c_i)\\ &= -\gamma\beta+\sum_{A\subseteq C, A\neq\emptyset}\gamma\cdot m_\mu(A)\cdot\min_{c_i\in A} g_{\boldsymbol{x}}(c_i). \qquad (30)\end{aligned}$$

Let $\psi : 2^C \to \mathbb{R}$ and $\phi_{\boldsymbol{x}} : 2^C \to \mathbb{R}$ be the mappings defined, respectively, as

$$\begin{aligned}\psi(\emptyset) &= -\gamma\beta,\\ \psi(A) &= \gamma\cdot m_\mu(A), \quad \forall A\subseteq C, A\neq\emptyset,\end{aligned}$$

and

$$\begin{aligned}\phi_{\boldsymbol{x}}(\emptyset) &= 1,\\ \phi_{\boldsymbol{x}}(A) &= \min_{c_i\in A} g_{\boldsymbol{x}}(c_i), \quad \forall A\subseteq C, A\neq\emptyset.\end{aligned}$$

Then, from (30), we obtain

$$\begin{aligned}\gamma\left(\mathcal{C}_\mu(g_{\boldsymbol{x}})-\beta\right) &= \sum_{A\subseteq C}\psi(A)\cdot\phi_{\boldsymbol{x}}(A)\\ &= \boldsymbol{\psi}^\intercal\boldsymbol{\phi}_{\boldsymbol{x}}, \qquad (31)\end{aligned}$$

with $\boldsymbol{\psi}$ and $\boldsymbol{\phi}_{\boldsymbol{x}}$ the column vectors whose elements are, respectively, the values $\psi(A)$ and $\phi_{\boldsymbol{x}}(A)$, $A\subseteq C$, ordered according to the same (arbitrary) order on the subsets of C.

Let us note that the original parameters $\boldsymbol{\upsilon} = (\gamma, \boldsymbol{m}_\mu, \beta)$ of CR can be recovered from the alternative parameters $\boldsymbol{\psi}$ as follows:

$$\begin{aligned}\gamma &= \sum_{A\subseteq C, A\neq\emptyset}\psi(A),\\ \beta &= \frac{-\psi(\emptyset)}{\sum_{A\subseteq C, A\neq\emptyset}\psi(A)},\\ m_\mu(A) &= \frac{\psi(A)}{\sum_{B\subseteq C, B\neq\emptyset}\psi(B)}, \quad \forall A\subseteq C, A\neq\emptyset,\end{aligned}$$

using the fact that $\mu(C) = 1 = \sum_{B\subseteq C, B\neq\emptyset} m_\mu(B)$.

Under this reparametrisation, fitting the CR model amounts to determining the actual values $\hat{\boldsymbol{\psi}}$ of the parameters $\boldsymbol{\psi}$. The conditional likelihood of these alternative parameters is given by

$$L_{\mathcal{D}}(\boldsymbol{\psi}) = \prod_{i=1}^{n}\left(h(\boldsymbol{\psi}^\intercal\boldsymbol{\phi}_{\boldsymbol{x}^{(i)}})\right)^{y^{(i)}}\cdot\left(1-h(\boldsymbol{\psi}^\intercal\boldsymbol{\phi}_{\boldsymbol{x}^{(i)}})\right)^{(1-y^{(i)})}, \qquad (32)$$

and the log-likelihood is then

$$\ell_{\mathcal{D}}(\boldsymbol{\psi}) = \sum_{i=1}^{n} y^{(i)} \log(h(\boldsymbol{\psi}^{\intercal}\boldsymbol{\phi}_{\boldsymbol{x}^{(i)}})) + (1 - y^{(i)}) \log(1 - h(\boldsymbol{\psi}^{\intercal}\boldsymbol{\phi}_{\boldsymbol{x}^{(i)}})).$$

Given the constraints (14) and (15) as well as the relations between parameters $\boldsymbol{\upsilon}$ and $\boldsymbol{\psi}$, the values $\hat{\boldsymbol{\psi}}$ are obtained by solving the following constrained optimisation problem:

$$\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi} \in \Psi} \ell_{\mathcal{D}}(\boldsymbol{\psi}) \tag{33}$$

with Ψ the set composed of the vectors $\boldsymbol{\psi} \in \mathbb{R}^{2^m}$ satisfying

$$\sum_{A \subseteq C, A \neq \emptyset} \psi(A) \geq -\psi(\emptyset) \geq 0, \qquad \sum_{A \subseteq C, A \neq \emptyset} \psi(A) > 0,$$

and

$$\sum_{B \subseteq A \setminus \{c_i\}} \psi(B \cup \{c_i\}) \geq 0, \quad \forall A \subseteq C,\ \forall c_i \in C.$$

In other words, and as already remarked in [17, Section 6], CR can be seen as fitting a (constrained) linear function in the feature space spanned by the set of features $\{\phi_{\boldsymbol{x}}(A) | A \subseteq C\}$. From a formal point of view, this reparametrisation presents thus CR under a similar form to that of LR. This is particularly interesting as it makes it possible to extend straightforwardly to CR both the approach followed by Minary *et al.* [31] to derive ELR and some other useful results related to LR, as will be seen in the next section. Furthermore, from a practical of view, let us remark that we have observed (for instance, when running the experiments reported in Section 5) that solving the optimisation problem (33) is on average an order of magnitude faster than solving the optimisation problem (13) when using the `fmincon` function of Matlab (function reportedly used in [17] to solve (13)).

4.2 The evidential choquistic regression model

In this section, an evidential extension of CR, called evidential choquistic regression (ECR), is derived. This extension may be obtained equivalently using the CR model based on the expression (8) or based on the linear form (31). Since this latter expression has the advantages mentioned in the preceding section, it is preferred.

Similarly to the ELR model recalled in Section 3.3, we propose to see the label $y \in \mathcal{Y} = \{0, 1\}$ of an object whose feature vector $\boldsymbol{x}$ has been observed, as the realisation of a binary random variable Y following a Bernoulli distribution with parameter $\theta = h(\boldsymbol{\psi}^{\intercal}\boldsymbol{\phi}_{\boldsymbol{x}})$, where there is some uncertainty on $\boldsymbol{\psi}$ induced by the observation of the training data $\mathcal{D}$. This latter uncertainty is represented by the likelihood-based (consonant) belief function $Bel^{\Psi}_{\mathcal{D}}$ whose contour function is defined by

$$pl^{\Psi}_{\mathcal{D}}(\boldsymbol{\psi}) = \frac{L_{\mathcal{D}}(\boldsymbol{\psi})}{L_{\mathcal{D}}(\hat{\boldsymbol{\psi}})}, \quad \forall \boldsymbol{\psi} \in \Psi,$$

and whose corresponding plausibility function is obtained as

$$Pl^{\Psi}_{\mathcal{D}}(A) = \sup_{\boldsymbol{\psi} \in A} pl^{\psi}_{\mathcal{D}}(\boldsymbol{\psi}).$$

Then, the uncertainty with respect to θ induced by the uncertainty on $\boldsymbol{\psi}$ and the observed feature vector $\boldsymbol{x}$, is represented by the consonant belief function $Bel^{\Theta}_{\mathcal{D},\boldsymbol{x}}$ whose contour function is given by

$$pl^{\Theta}_{\mathcal{D},\boldsymbol{x}}(\theta) = \begin{cases} 0 & \text{if } \theta \in \{0, 1\}, \\ Pl^{\Psi}_{\mathcal{D}}(\{\boldsymbol{\psi} \in \Psi \mid \theta = h(\boldsymbol{\psi}^{\intercal}\boldsymbol{\phi}_{\boldsymbol{x}})\}) & \text{otherwise,} \end{cases}$$

where

$$\begin{aligned}\{\boldsymbol{\psi} \in \Psi \mid \theta = h(\boldsymbol{\psi}^\intercal \boldsymbol{\phi}_{\boldsymbol{x}})\} &= \left\{\boldsymbol{\psi} \in \Psi \mid -\boldsymbol{\psi}^\intercal \boldsymbol{\phi}_{\boldsymbol{x}} = \ln(\theta^{-1} - 1)\right\} \\ &= \left\{\boldsymbol{\psi} \in \Psi \mid \psi(\emptyset) = -\ln(\theta^{-1} - 1) - \sum_{A \subseteq C, A \neq \emptyset} \psi(A) \cdot \phi_{\boldsymbol{x}}(A)\right\},\end{aligned}$$

which yields

$$pl^{\Theta}_{\mathcal{D},\boldsymbol{x}}(\theta) = \sup_{\boldsymbol{\psi}^* \in \mathbb{R}^{2^m - 1}} pl^{\Psi}_{\mathcal{D}}\left(-\ln(\theta^{-1} - 1) - \sum_{A \subseteq C, A \neq \emptyset} \psi(A) \cdot \phi_{\boldsymbol{x}}(A), \boldsymbol{\psi}^*\right), \quad \forall \theta \in (0, 1), \tag{34}$$

with $\boldsymbol{\psi}^*$ the vector whose elements are the values $\psi(A)$, $A \subseteq C, A \neq \emptyset$. The value $pl^{\Theta}_{\mathcal{D},\boldsymbol{x}}(\theta)$, $\theta \in (0, 1)$, can be obtained by an iterative maximisation algorithm.

Similarly as for ELR, the prediction of Y given this uncertainty on θ is carried out using the approach for prediction recalled in Section 3.2. Specifically, the predictive belief $Bel^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{1\})$ and predictive plausibility $Pl^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{1\})$ of the positive class given the predictors are obtained from the contour function $pl^{\Theta}_{\mathcal{D},\boldsymbol{x}}$ defined by (34) using Eqs. (26) and (27), respectively.

Example 6. *Let us illustrate ECR on the classification problem with training dataset $\mathcal{D}$ of Example 1. The contour function $pl^{\Theta}_{\mathcal{D},\boldsymbol{x}}(\theta)$ obtained from* (34) *for feature vector $\boldsymbol{x} = (2, -1)^\intercal$ is illustrated in Figure 6a. This contour function induces*

$$Bel^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{1\}) = 0.3403, \quad Pl^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{1\}) = 0.8889.$$

Equivalently, we have

$$m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{0\}) = 0.1111, \quad m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\{1\}) = 0.3403, \quad m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\mathcal{Y}) = 0.5485.$$

Using the interval dominance decision rule with $m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}$ and 0/1 loss, the decision reached for the label y of the object with feature vector $\boldsymbol{x} = (2, -1)^\intercal$ is $y = \{0, 1\}$. In other words, the uncertainty is too high and an imprecise decision is obtained when using the evidential extension of CR (CR yielded the decision $y = 1$, see Example 2). This contrasts with LR where the same precise decision ($y = 1$) was obtained with its evidential extension (see Examples 1 and 5, respectively).

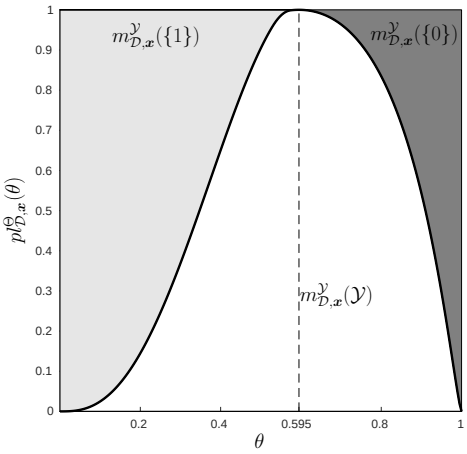
For completeness, the decision regions (in the case of 0/1 loss) and the levels of ignorance of ECR are shown in Figures 6b and 6c, respectively. Similarly as for ELR, regions where imprecise decisions are made correspond to areas with greatest levels of ignorance.

Remark 1. *In order to prevent LR from over-fitting when the training examples in $\mathcal{D}$ are perfectly separable, a modified version due to Platt [38] of the conditional likelihood $L_{\mathcal{D}}$ of the LR parameters may be used: the label $y^{(i)} \in \{0, 1\}$ in* (3) *is then replaced by $t^{(i)} \in [0, 1]$ defined by*

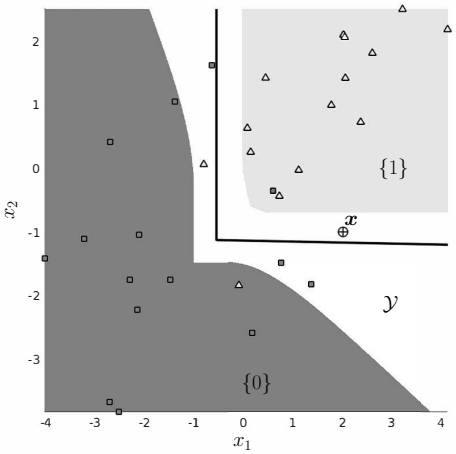
$$t^{(i)} = \begin{cases} \frac{N_+ + 1}{N_+ + 2} & \text{if } y^{(i)} = 1, \\ \frac{1}{N_- + 2} & \text{if } y^{(i)} = 0, \end{cases} \tag{35}$$

where N_+ and N_- are the number of positive and negative samples, respectively, in $\mathcal{D}$. This modification ensures $L_{\mathcal{D}}$ to have a unique supremum $\hat{\boldsymbol{\sigma}}$. In their derivation of ELR – precisely, in their definition (28) *of the contour function $pl^{\Sigma}_{\mathcal{D}}$ – Minary* et al. *[31] actually used this modified version of $L_{\mathcal{D}}$.*

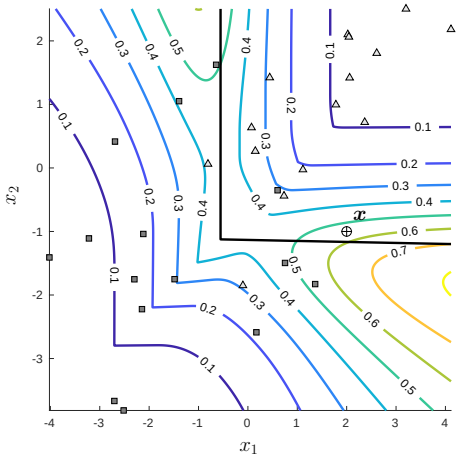
As already mentioned in Section 2.3, CR may also be prone to over-fitting. Besides restricting μ to be k-additive for some $k < m$, it is proposed in [17] to address this issue by adding a L_1-regulariser on the Möbius transform in the objective (13)*. However, we may as well consider a modified version of the conditional likelihood $L_{\mathcal{D}}$ of the CR parameters, where the label $y^{(i)}$ in* (11) *(or in* (32) *if the reparametrisation is used) has been replaced by $t^{(i)}$ defined by* (35)*. Let us remark that, thanks to the reparameterised form of CR and its formal similarity with LR, it is then straightforward to obtain that this latter modification ensures $L_{\mathcal{D}}$ to have a unique supremum $\hat{\boldsymbol{\psi}}$.*



(a) Contour function $pl^{\Theta}_{\mathcal{D},\boldsymbol{x}}(\theta)$ for $\boldsymbol{x} = (2, -1)$.



(b) Training instances $\mathcal{D}$. Object of interest ($\oplus$). Decision boundary ($-$) of CR. Decision regions of ECR according to the interval dominance decision rule: $y = 1$ (light grey), $y = 0$ (dark grey) and $y \in \{0, 1\}$ (white).



(c) Training instances $\mathcal{D}$. Decision boundary ($-$) of CR. Contour lines of the ignorance $m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\mathcal{Y})$ obtained with ECR.

Figure 6: Evidential choquistic regression.

In the experiments reported in Section 5, the modified labels (35) *(rather than the original labels* $y^{(i)}$*) are not only used in the conditional likelihood involved in ELR (as it is the actual definition of ELR given in [31]), but they are also used in all the conditional likelihoods involved in LR, CR and ECR. This is done in order to ensure a fairer comparison between the approaches.*

4.3 Complexity

In this section, we provide a brief analysis of the computational complexity involved in making a prediction using ECR. To derive this complexity, we start by recalling the complexity of training LR, then we proceed with the complexity of making a prediction using ELR and with that of training CR, before ending with that of making a prediction using ECR.

Given a training set $\mathcal{D}$ composed of n instances described by m-dimensional feature vectors, the computational complexity of finding the MLE of LR according to Eq. (4), using gradient ascent (GA), is [32]:

$$\mathcal{O}\left((m+1)\cdot n\cdot t\right),$$

with t the number of ascent iterations.

Making a prediction using ELR involves essentially computing contour function $pl^{\Theta}_{\mathcal{D},\boldsymbol{x}}$ (inducing the predictive belief function $Bel^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}$), which in practice is approximated by: (1) selecting a finite number c of values of $\theta \in (0,1)$; (2) evaluating $pl^{\Theta}_{\mathcal{D},\boldsymbol{x}}$, that is, solving the optimisation problem (29), for each selected value of θ; (3) interpolating $pl^{\Theta}_{\mathcal{D},\boldsymbol{x}}$ linearly between the selected values of θ. Note that the same approach is followed for making a prediction using ECR (in step (2), optimisation problem (34) is then solved instead of (29)).

The complexity of solving the optimisation problem (29), using GA, is similar to that of finding the MLE of LR: it is $\mathcal{O}\left(m\cdot n\cdot t\right)$. Since this optimisation is done c times, the complexity of making a prediction using ELR is

$$\mathcal{O}\left(c\cdot m\cdot n\cdot t\right).$$

Training CR amounts to solving the constrained optimisation problem (33), which can be done using projected gradient ascent [2]. From the complexity of training LR, we obtain that the complexity of training CR is

$$\mathcal{O}\left(2^m\cdot n\cdot t\right).$$

We thus face an exponential complexity in the number of features. This complexity can nonetheless be significantly reduced by restricting the capacity underlying CR to be k-additive, for some $k < m$, which, we recall, is also useful to prevent overfitting. Using this restriction, the dimension of the parameter space Ψ of CR is reduced to $\sum_{l=0}^{k}\binom{m}{l}$, and the complexity of training CR becomes

$$\mathcal{O}\left(\sum_{l=0}^{k}\binom{m}{l}\cdot n\cdot t\right).$$

As already mentioned, making a prediction using ECR follows the same approach to that of making a prediction using ELR. We have in particular that ECR inherits the complexity of CR, in the same manner as ELR inherits the complexity of LR. Therefore, the complexity of making a prediction using ECR is

$$\mathcal{O}\left(c\cdot 2^m\cdot n\cdot t\right).$$

However, using the restriction to k-additive capacities as done in the experiments reported in Section 5, the complexity comes down to

$$\mathcal{O}\left(c\cdot\sum_{l=1}^{k}\binom{m}{l}\cdot n\cdot t\right),$$

which remains acceptable for real applications having low dimensional features and for k small.

Name	#instances	#features	source
Blood Transfusion Service Center (BLO)	748	4	UCI
Contraceptive Method Choice (CMC)	1473	9	UCI
Haberman's Survival (HAB)	306	3	UCI
Hamster (HAM)	73	5	StatLib
Lecturers Evaluation (LEV)	1000	4	WEKA
Mammographic (MMG)	961	4	UCI
Auto MPG (MPG)	398	7	UCI
Teaching Assistant Evaluation (TAE)	151	5	UCI
Yeast (YST)	1484	8	UCI

Table 1: Datasets used in the experiments

5 Experiments

In this section, the results of some experiments that we conducted are reported. Their goal is twofold. First, we would like to show that by capturing uncertainty in a more subtle way, ECR is a useful reliable variant of CR. Secondly, we seek to show that by enabling more flexibility with respect to the possibility of interactions between the predictors, ECR is competitive with ELR. Our experiments rely on some (monotone) datasets, which are presented in the next section.

5.1 Datasets

We used nine datasets, summarised in Table 1, for which the assumption of monotonicity in the input variables seems reasonable. To our knowledge, at least some of them have also been used in previous works on monotonic classification, such as in [18, 17, 4]. They are available from the UCI[1], StatLib[2] and WEKA[3] dataset repositories. Some of them have numerical or categorical outputs, which were thus binarised. Furthermore, all the input features were normalised using (10). The following paragraphs provide some more details on each one of the datasets.

Blood Transfusion Service Center (BLO) This dataset is a subsample of the database maintained by the Blood Transfusion Service Center in Hsin-Chu City in Taiwan. The data represents 748 donor examples described by 4 features : recency (months since last donation), frequency (total number of donation), monetary (total blood donated), time (months since first donation). The goal is to predict the dependent binary variable representing whether the donor donated blood in March 2007.

Contraceptive Method Choice (CMC) This dataset is composed of 1473 instances collected in 1987 by the National Indonesia Contraceptive Prevalence Survey. The problem is to predict the choice of contraceptive method encoded with 3 categories (no use, long-term methods, or short-term methods) from 9 demographic and socio-economic features of women: age, education, education of husband, number of children, religious, working, occupation of husband, living index, media exposure. The categorical output was binarised into a two-class target by distinguishing between those women that do not use a contraceptive method (category "no use") and those that do use one (categories "long-term methods" and "short-term methods").

Haberman's Survival (HAB) This dataset contains 306 instances on the survival of patients who had undergone surgery for breast cancer between 1958 and 1970 at the University of Chicago's Billings Hospital. Instances are described by 3 ordinal features: patient's age, year of operation, and amount of

[1] https://archive.ics.uci.edu
[2] http://lib.stat.cmu.edu/datasets
[3] https://waikato.github.io/weka-wiki/datasets

detected axillary nodes. The goal is to predict the dependent binary variable representing whether the patient died within 5 years.

Hamster (HAM) This dataset contains the measurements of the weights of 6 organs (lung, heart, liver, spleen, kidney, testes) of 73 hamsters from a strain with a congenital heart problem. The numeric output is the weight measurement of testes, which was binarised as done in the OpenML dataset repository[4], *i.e.*, it was binarised to a two-class target by computing the mean and classifying all instances with a measurement lower than the mean as positive and all others as negative.

Lecturers Evaluation (LEV) This dataset is composed of examples of anonymous lecturer evaluations sampled at the end of courses where students were asked to score lecturers according to four ordinal features including oral skills and contribution to their professional/general knowledge. The ordinal output is an overall evaluation of a lecturer's performance which was binarised by interpreting the scores between 3 and 4 as positive (good evaluation) and the scores between 0 and 2 as negative (bad evaluation), as done in [17].

Mammographic (MMG) This dataset is composed of 961 instances of patients. The problem is to predict whether a mammographic mass lesion is benign or malignant, given 3 BI-RADS features (mass shape, mass margin, density) and the patient's age.

Auto MPG (MPG) This dataset was designed for the American Statistical Association Exposition of 1983. The problem is about the prediction of the city-cycle fuel consumption (in miles per gallon) based on seven features of a car: cylinders, displacement, horsepower, weight, acceleration, model year, origin. The numerical consumption output was binarised by thresholding at the median, as done in [17]. Furthermore, incomplete instances were removed.

Teacher Assistant Evaluation (TAE) The data was collected at the Statistics Department of the University of Wisconsin-Madison. It represents a collection of 151 evaluations of teacher assistants based on 6 ordinal features: native speaker, instructor, course, semester, class size. The ordinal output represents 3 categories (low, medium, and high), which was binarised to a two-class target by distinguishing between great (category high) and not great evaluation (categories medium and low).

Yeast (YST) This dataset is composed of 1484 instances describing the cellular localisation sites of proteins, given 8 features extracted from the amino acid contents. The protein localisation site is a nominal output encoded with 10 modalities ("cyt" for cytosolic or cytoskeletal, "nuc" for nuclear, "mit" for mitochondrial,...). It was converted to a two-class target by relabelling the majority localisation site modality as positive and all others as negative, which amounts to predicting whether the localisation site is "cyt".

5.2 Accuracy

As stressed in [43], it is important to make sure that an improved representation of uncertainty permitted by a reliable variant of a classifier, does not come at the expense of a loss in accuracy. Therefore, we looked first at the predictive accuracy of our reliable version of CR, *i.e.*, of ECR, before investigating its potential benefits with respect to CR in terms of uncertainty representation. This also allowed us to check whether the superior accuracy of CR in comparison to LR observed in [17] when the amount of training data is sufficiently large[5], is carried over to their evidential extensions, as this seems to be another important requirement.

[4]See https://www.openml.org/d/893

[5]In [17], it was observed that CR outperforms LR when there is sufficiently extensive training data, but also that LR may be better if the amount of training data is small, which the authors explained by the fact that the flexibility of CR, *i.e.*, its ability to capture nonlinear dependencies between input attributes, may lead to overfitting when there is few data and is becoming more and more advantageous as more data is available.

For each dataset, we followed a similar procedure to that of [17]. The dataset was randomly split into two parts: a training set and a test set. Each of the models (LR, CR, ELR, ECR) was learnt on the training set and its performance was measured using the standard misclassification rate (0/1 loss) estimated on the test set – for the reliable classifiers ELR and ECR, a (precise) prediction was obtained using the decision rule (20). This process was repeated 100 times and the results were averaged over these repetitions. Furthermore, we considered three proportions between training and test data, which were 20:80, 50:50 and 80:20, in order to study the influence of the amount of training data. In addition, to obtain the best accuracy, the underlying capacity of CR (and thus of ECR, as it extends CR) was restricted to be k-additive, with k determined by 10-fold cross validation on the training set.

The results for this experiment are given in Tables 2 and 3. First, as can be seen from Table 2, ECR and ELR have essentially the same predictive accuracies as CR and LR, respectively. Second, ECR has a better predictive accuracy than ELR (this is all the more true as the training set size increases).

In order to further confirm these observed differences in the classifier performances, we resorted to the statistical tests recommended by Demšar [9]. First, a Friedman test (using the F_F statistic, see [9, Section 3.2.2]) was used for comparing LR, CR, ELR and ECR over the nine datasets, with the null hypothesis that the classifiers are equivalent. The test statistic relies on the average rankings of the classifiers (provided in Table 2), where the rankings are based in this experiment on the misclassification rate. We found that the null hypothesis is rejected at significance level $\alpha = 0.05$, for any of the three proportions considered between training and test data. When the null hypothesis is rejected, the classifiers can then be compared to each other using a post-hoc test such as the one of Bonferroni-Dunn [9], which relies also on the average rankings of the classifiers. We used this test to compare ECR to the other classifiers at significance level $\alpha = 0.05$. We found that ECR is significantly better than both LR and ELR and that the difference with CR is not significant, for any of the three proportions considered between training and test data.

Overall, these results mean that the two requirements set above for ECR seem thus to be met: enabling the possibility to capture extra information on the reliability of predictions does not deteriorate the predictive accuracy, and enabling nonlinear dependencies between input attributes brings an improvement in predictive accuracy not only in the probabilistic setting but also in the evidential one. Moreover, this latter flexibility in terms of dependencies between input attributes is all the more exploited to improve accuracy as the learning set size increases, as can be seen from Table 3, which shows that the optimal (cross-validated) value of k is increasing in the learning set size.

5.3 Utility-discounted accuracy

As it was illustrated by Examples 5 and 6, an interest of the improved representation of uncertainty offered by the reliable variants of LR and CR is the ability to make imprecise predictions when there is too much uncertainty. The question of the measurement of the quality of classifiers issuing potentially imprecise predictions has received a few answers (see [51, 50] and the references therein). In general, the idea is to have a measure taking both into account the accuracy and the precision of the predictions. A standard approach is to extend the classical accuracy measure to imprecise predictions, which makes it then possible to compare classifiers issuing imprecise predictions with classifiers issuing only precise predictions.

A first proposal in this vein, known as discounted accuracy, reads as follows. Let $\mathcal{T} = \{\boldsymbol{x}^{(i)}, y^{(i)}\}_{i=1}^{T}$ be a test set composed of T objects, with $y^{(i)} \in \mathcal{Y}$ the known (and precise) label of object i. Furthermore, let $A^{(i)} \subseteq \mathcal{Y}$ denote the (potentially imprecise) prediction for the label of object i, provided by a classifier having observed feature vector $\boldsymbol{x}^{(i)}$. Then, the discounted accuracy Acc_D of this classifier is:

$$\mathrm{Acc}_D = \frac{1}{T} \sum_{i=1}^{T} \mathbf{1}_{A^{(i)}}(y^{(i)}) \cdot \frac{1}{|A^{(i)}|}, \tag{36}$$

with $\mathbf{1}_{A^{(i)}}$ the indicator function of subset $A^{(i)}$. This measure clearly degenerates into the classical accuracy measure when all $A^{(i)}$, $i = 1, \ldots, T$, are singletons, *i.e.*, all the predictions are precise.

An issue with the discounted accuracy is that the value of an imprecise prediction $A^{(i)}$ is considered to be the same as that of a purely random choice within $A^{(i)}$, as criticised by Zaffalon *et al.* [51]. In

Datasets	CR	ECR	LR	ELR
BLO(20:80)	.2256 ± .0125(2)	.2255 ± .0126(1)	.2271 ± .0128(4)	.2269 ± .0125(3)
CMC(20:80)	.3190 ± .0151(1)	.3217 ± .0221(2)	.3350 ± .0128(4)	.3349 ± .0129(3)
HAB(20:80)	.2607 ± .0184(1)	.2608 ± .0189(2)	.2611 ± .0210(3.5)	.2611 ± .0211(3.5)
HAM(20:80)	.4498 ± .0812(1)	.4512 ± .0796(2)	.4672 ± .0683(3.5)	.4672 ± .0680(3.5)
LEV(20:80)	.1543 ± .0116(1)	.1546 ± .0115(2)	.1628 ± .0098(3.5)	.1628 ± .0099(3.5)
MMG(20:80)	.2074 ± .0126(2)	.2066 ± .0126(1)	.2087 ± .0116(3)	.2088 ± .0116(4)
MPG(20:80)	.1093 ± .0168(2)	.1088 ± .0168(1)	.1102 ± .0173(4)	.1098 ± .0175(3)
TAE(20:80)	.3346 ± .0450(1)	.3371 ± .0480(2)	.3373 ± .0463(3)	.3376 ± .0474(4)
YST(20:80)	.3098 ± .0101(1)	.3216 ± .0553(4)	.3181 ± .0107(2)	.3183 ± .0105(3)
Average ranking	1.33	1.89	3.39	3.39
BLO(50:50)	.2225 ± .0165(1)	.2226 ± .0166(2)	.2259 ± .0162(3)	.2261 ± .0162(4)
CMC(50:50)	.3026 ± .0127(1)	.3040 ± .0128(2)	.3284 ± .0117(3)	.3285 ± .0117(4)
HAB(50:50)	.2537 ± .0278(2)	.2527 ± .0278(1)	.2556 ± .0284(4)	.2553 ± .0283(3)
HAM(50:50)	.3983 ± .0862(1)	.4047 ± .0781(2)	.4453 ± .0716(4)	.4444 ± .0717(3)
LEV(50:50)	.1423 ± .0118(1.5)	.1423 ± .0117(1.5)	.1637 ± .0121(3.5)	.1637 ± .0122(3.5)
MMG(50:50)	.1983 ± .0147(2)	.1982 ± .0148(1)	.1999 ± .0132(3)	.2000 ± .0132(4)
MPG(50:50)	.0910 ± .0204(2)	.0909 ± .0202(1)	.0978 ± .0178(4)	.0977 ± .0177(3)
TAE(50:50)	.3139 ± .0468(1)	.3141 ± .0476(2)	.3267 ± .0432(4)	.3264 ± .0432(3)
YST(50:50)	.3100 ± .0131(1)	.3101 ± .0130(2)	.3211 ± .0140(4)	.3210 ± .0140(3)
Average ranking	1.39	1.61	3.61	3.39
BLO(80:20)	.2195 ± .0291(1.5)	.2195 ± .0293(1.5)	.2307 ± .0283(4)	.2306 ± .0282(3)
CMC(80:20)	.2968 ± .0215(1)	.3005 ± .0220(2)	.3245 ± .0250(3.5)	.3245 ± .0250(3.5)
HAB(80:20)	.2548 ± .0508(1)	.2549 ± .0507(2)	.2582 ± .0500(4)	.2579 ± .0505(3)
HAM(80:20)	.3413 ± .1110(1)	.3507 ± .1155(2)	.4360 ± .1061(4)	.4353 ± .1040(3)
LEV(80:20)	.1429 ± .0220(1.5)	.1429 ± .0222(1.5)	.1657 ± .0205(3.5)	.1657 ± .0205(3.5)
MMG(80:20)	.1910 ± .0243(1.5)	.1910 ± .0244(1.5)	.1949 ± .0236(3.5)	.1949 ± .0236(3.5)
MPG(80:20)	.0838 ± .0288(2)	.0836 ± .0281(1)	.0905 ± .0295(3.5)	.0905 ± .0297(3.5)
TAE(80:20)	.3050 ± .0814(2)	.3043 ± .0813(1)	.3337 ± .0788(4)	.3323 ± .0803(3)
YST(80:20)	.3095 ± .0221(1)	.3102 ± .0221(2)	.3200 ± .0230(3.5)	.3200 ± .0230(3.5)
Average ranking	1.39	1.61	3.72	3.28

Table 2: Mean and standard deviation of 0/1 loss (associated ranking) for CR, LR and their reliable variants, trained on 20%, 50%, and 80% of the datasets.

Datasets	k
BLO(20:80)	1.33 ± 0.49
CMC(20:80)	2.90 ± 1.48
HAB(20:80)	1.52 ± 0.66
HAM(20:80)	2.16 ± 0.98
LEV(20:80)	2.29 ± 1.05
MMG(20:80)	1.96 ± 0.95
MPG(20:80)	2.17 ± 1.13
TAE(20:80)	2.07 ± 1.00
YST(20:80)	2.78 ± 1.28
BLO(50:50)	1.70 ± 0.46
CMC(50:50)	3.42 ± 1.51
HAB(50:50)	1.89 ± 0.75
HAM(50:50)	2.65 ± 0.89
LEV(50:50)	3.29 ± 0.57
MMG(50:50)	2.17 ± 0.91
MPG(50:50)	2.74 ± 1.19
TAE(50:50)	2.47 ± 1.12
YST(50:50)	3.34 ± 1.25
BLO(80:20)	1.97 ± 0.17
CMC(80:20)	3.85 ± 1.69
HAB(80:20)	2.00 ± 0.75
HAM(80:20)	3.11 ± 0.63
LEV(80:20)	3.55 ± 0.50
MMG(80:20)	2.16 ± 0.77
MPG(80:20)	3.00 ± 1.03
TAE(80:20)	3.04 ± 1.00
YST(80:20)	3.70 ± 0.96

Table 3: Mean (± standard deviation) value of k determined by cross-validation over the 100 repetitions.

other words, in the case of ambiguity, it does not reward caution. To address this problem, Zaffalon *et al.* [51] justified another measure called utility-discounted accuracy, which also degenerates to the classical accuracy measure when all the predictions are precise. This measure is defined as follows:

$$\mathrm{Acc}_U = \frac{1}{T}\sum_{i=1}^{T} \mathbf{1}_{A^{(i)}}(y^{(i)}) \cdot u(\frac{1}{|A^{(i)}|}), \tag{37}$$

where u is a concave function on $[0,1]$, such that $u(1) = 1$ and $u(0) = 0$. Function u is interpreted as modelling the utility of a prediction according to its precision (the more precise the prediction, the higher the utility). In particular, it rewards an imprecise prediction $A^{(i)} \subseteq \mathcal{Y}$ at least as much as the discounted accuracy since this latter accuracy corresponds to giving a utility $\frac{1}{|A^{(i)}|}$ to this prediction, whereas prediction $A^{(i)}$ gets a utility $u(\frac{1}{|A^{(i)}|}) \geq \frac{1}{|A^{(i)}|}$ thanks to the concavity of u.

In the case of binary classification ($\mathcal{Y} = \{0,1\}$), which is of interest in this paper, Zaffalon *et al.* [51] proposed to reward the imprecise prediction $A^{(i)} = \{0,1\}$ by a utility between 0.65 and 0.8, *i.e.*, we should have $u(\frac{1}{|\{0,1\}|}) = u(0.5) \geq 0.65$ and $u(0.5) \leq 0.8$ (in contrast to the discounted accuracy, which corresponds to giving a utility of only 0.5 to such prediction). This means that by setting $u(0.5) = 0.65$, one is modelling a moderate cautiousness-seeking attitude – imprecise predictions are valued (rewarded), yet to a small extent – whereas by setting $u(0.5) = 0.80$, imprecise predictions are rewarded to a greater extent, hence modelling a stronger cautiousness-seeking attitude.

With that in mind, we repeated the experimental procedure described in Section 5.2, but with two modifications. First, instead of using the decision rule (20) for ELR and ECR, we use the interval dominance decision rule (21), therefore yielding potentially imprecise decisions for these classifiers. Second, instead of measuring the performances of LR, CR, ELR and ECR using the misclassification rate, we used the utility-discounted accuracy[6] with $u(0.5) = 0.65$ and also with $u(0.5) = 0.8$; the former performance measure is denoted by $\mathrm{Acc}_{U_{65}}$ and the latter is denoted by $\mathrm{Acc}_{U_{80}}$.

The results of this second experiment are presented in Tables 4 and 5, respectively, for the performance measures $\mathrm{Acc}_{U_{65}}$ and $\mathrm{Acc}_{U_{80}}$, respectively. As can be seen from these tables, ECR and ELR have globally better performances than CR and LR, respectively. Furthermore, the difference in performance between a classifier and its reliable version is generally all the greater as there are fewer training data. Overall, this suggests that ECR and ELR are able to capture useful information on the reliability of predictions and, in particular, that the greater uncertainty induced by fewer training data is appropriately taken into account by these reliable variants of CR and LR. Comparing the performances of ECR and ELR, we can see that ECR is globally better than ELR. In addition, similarly as in the case of predictive accuracy (see Section 5.2), the difference in performance between ECR and ELR is typically all the greater as there are more training data. This means that the flexibility of ECR is beneficial also when making reliable predictions and is all the more so as the learning set size increases.

Similarly as in Section 5.2, Friedman tests based on the average rankings (provided in Tables 4 and 5) induced by performance measures $\mathrm{Acc}_{U_{65}}$ and $\mathrm{Acc}_{U_{80}}$, were used for comparing further the classifiers. We found for both of these performance measures that the null hypothesis (the classifiers are equivalent) is rejected at level $\alpha = 0.05$, for any of the three proportions considered between training and test data. We proceeded with Bonferroni-Dunn tests (with level $\alpha = 0.05$) based on the average rankings induced by $\mathrm{Acc}_{U_{65}}$ and $\mathrm{Acc}_{U_{80}}$. We found that according to $\mathrm{Acc}_{U_{65}}$, ECR is significantly better than LR and that the differences with CR and ELR are not significant, whereas according to $\mathrm{Acc}_{U_{80}}$, ECR is significantly better than both LR and CR and the difference with ELR is not significant; all of these statements are obtained for any of the three proportions considered between training and test data. The non-significant differences with ELR according to these particular tests, mitigate the conclusion drawn above from Tables 4 and 5 that ECR is globally better than ELR when making reliable predictions.

To sum up this section and the previous one, we can conclude that ECR is a useful extension of both CR and ELR. It combines the advantages of both in terms of modelling capabilities, which is reflected by the results of our experiments. More precisely, the ability of ECR to capture uncertainty in

[6] Obviously, for LR and CR, this measure is nothing but their predictive accuracy (already provided in Section 5.2 in the form of the misclassification rate), as these models yield only precise predictions.

Datasets	CR	ECR	LR	ELR
BLO(20:80)	.7744 ± .0125(3)	.7848 ± .0117(1)	.7729 ± .0128(4)	.7838 ± .0108(2)
CMC(20:80)	.6810 ± .0151(3)	.7115 ± .0113(1)	.6650 ± .0128(4)	.6966 ± .0094(2)
HAB(20:80)	.7393 ± .0184(3)	.7528 ± .0163(1)	.7389 ± .0210(4)	.7513 ± .0163(2)
HAM(20:80)	.5502 ± .0812(3)	.6318 ± .0387(2)	.5328 ± .0683(4)	.6342 ± .0362(1)
LEV(20:80)	.8457 ± .0116(3)	.8555 ± .0086(1)	.8372 ± .0098(4)	.8542 ± .0089(2)
MMG(20:80)	.7926 ± .0126(3)	.8026 ± .0073(2)	.7913 ± .0116(4)	.8028 ± .0080(1)
MPG(20:80)	.8907 ± .0168(1)	.8722 ± .0224(4)	.8898 ± .0173(2)	.8866 ± .0118(3)
TAE(20:80)	.6654 ± .0450(3)	.6850 ± .0288(1)	.6627 ± .0463(4)	.6837 ± .0243(2)
YST(20:80)	.6902 ± .0101(3)	.7202 ± .0144(1)	.6819 ± .0107(4)	.7170 ± .0089(2)
Average ranking	2.78	1.56	3.78	1.89
BLO(50:50)	.7775 ± .0165(3)	.7856 ± .0167(1)	.7741 ± .0162(4)	.7852 ± .0167(2)
CMC(50:50)	.6974 ± .0127(2)	.7228 ± .0092(1)	.6716 ± .0117(4)	.6968 ± .0114(3)
HAB(50:50)	.7463 ± .0278(3)	.7587 ± .0251(1)	.7444 ± .0284(4)	.7567 ± .0255(2)
HAM(50:50)	.6017 ± .0862(3)	.6515 ± .0506(1)	.5547 ± .0716(4)	.6292 ± .0399(2)
LEV(50:50)	.8577 ± .0118(2)	.8637 ± .0094(1)	.8363 ± .0121(4)	.8523 ± .0109(3)
MMG(50:50)	.8017 ± .0147(3)	.8071 ± .0113(1)	.8001 ± .0132(4)	.8053 ± .0114(2)
MPG(50:50)	.9090 ± .0204(1)	.9055 ± .0151(2)	.9022 ± .0178(4)	.9029 ± .0152(3)
TAE(50:50)	.6861 ± .0468(3)	.7075 ± .0294(2)	.6733 ± .0432(4)	.7077 ± .0310(1)
YST(50:50)	.6900 ± .0131(3)	.7159 ± .0125(1)	.6789 ± .0140(4)	.7044 ± .0125(2)
Average ranking	2.56	1.22	4.00	2.22
BLO(80:20)	.7805 ± .0291(3)	.7827 ± .0277(2)	.7693 ± .0283(4)	.7834 ± .0295(1)
CMC(80:20)	.7032 ± .0215(2)	.7283 ± .0179(1)	.6755 ± .0250(4)	.6968 ± .0227(3)
HAB(80:20)	.7452 ± .0508(3)	.7567 ± .0502(1)	.7418 ± .0500(4)	.7512 ± .0488(2)
HAM(80:20)	.6587 ± .1110(2)	.6828 ± .0721(1)	.5640 ± .1061(4)	.6208 ± .0677(3)
LEV(80:20)	.8571 ± .0220(2)	.8651 ± .0191(1)	.8343 ± .0205(4)	.8496 ± .0182(3)
MMG(80:20)	.8090 ± .0243(2)	.8112 ± .0217(1)	.8051 ± .0236(4)	.8067 ± .0222(3)
MPG(80:20)	.9162 ± .0288(1)	.9156 ± .0235(2)	.9095 ± .0295(4)	.9114 ± .0269(3)
TAE(80:20)	.6950 ± .0814(3)	.7134 ± .0587(2)	.6663 ± .0788(4)	.7140 ± .0660(1)
YST(80:20)	.6905 ± .0221(3)	.7128 ± .0190(1)	.6800 ± .0230(4)	.7018 ± .0209(2)
Average ranking	2.33	1.33	4.00	2.33

Table 4: Mean and standard deviation of $\mathrm{Acc}_{U_{65}}$ (associated ranking) for CR, LR and their reliable variants, trained on 20%, 50%, and 80% of the datasets.

Datasets	CR	ECR	LR	ELR
BLO(20:80)	.7744 ± .0125(3)	.8058 ± .0196(1)	.7729 ± .0128(4)	.8045 ± .0192(2)
CMC(20:80)	.6810 ± .0151(3)	.7692 ± .0147(1)	.6650 ± .0128(4)	.7467 ± .0141(2)
HAB(20:80)	.7393 ± .0184(3)	.7833 ± .0198(1)	.7389 ± .0210(4)	.7823 ± .0202(2)
HAM(20:80)	.5502 ± .0812(3)	.7374 ± .0467(2)	.5328 ± .0683(4)	.7512 ± .0522(1)
LEV(20:80)	.8457 ± .0116(3)	.8822 ± .0116(1)	.8372 ± .0098(4)	.8765 ± .0117(2)
MMG(20:80)	.7926 ± .0126(3)	.8268 ± .0100(1)	.7913 ± .0116(4)	.8252 ± .0098(2)
MPG(20:80)	.8907 ± .0168(3)	.9125 ± .0139(2)	.8898 ± .0173(4)	.9196 ± .0108(1)
TAE(20:80)	.6654 ± .0450(3)	.7556 ± .0299(2)	.6627 ± .0463(4)	.7644 ± .0262(1)
YST(20:80)	.6902 ± .0101(3)	.7646 ± .0154(1)	.6819 ± .0107(4)	.7534 ± .0116(2)
Average ranking	3.00	1.33	4.00	1.67
BLO(50:50)	.7775 ± .0165(3)	.7965 ± .0184(1)	.7741 ± .0162(4)	.7945 ± .0183(2)
CMC(50:50)	.6974 ± .0127(3)	.7673 ± .0106(1)	.6716 ± .0117(4)	.7300 ± .0136(2)
HAB(50:50)	.7463 ± .0278(3)	.7748 ± .0270(1)	.7444 ± .0284(4)	.7722 ± .0273(2)
HAM(50:50)	.6017 ± .0862(3)	.7348 ± .0588(2)	.5547 ± .0716(4)	.7412 ± .0571(1)
LEV(50:50)	.8577 ± .0118(3)	.8805 ± .0096(1)	.8363 ± .0121(4)	.8673 ± .0119(2)
MMG(50:50)	.8017 ± .0147(3)	.8241 ± .0121(1)	.8001 ± .0132(4)	.8199 ± .0120(2)
MPG(50:50)	.9090 ± .0204(3)	.9286 ± .0138(1)	.9022 ± .0178(4)	.9219 ± .0152(2)
TAE(50:50)	.6861 ± .0468(3)	.7579 ± .0356(1)	.6733 ± .0432(4)	.7575 ± .0375(2)
YST(50:50)	.6900 ± .0131(3)	.7439 ± .0152(1)	.6789 ± .0140(4)	.7279 ± .0136(2)
Average ranking	3.00	1.11	4.00	1.89
BLO(80:20)	.7805 ± .0291(3)	.7912 ± .0284(1)	.7693 ± .0283(4)	.7906 ± .0305(2)
CMC(80:20)	.7032 ± .0215(3)	.7673 ± .0194(1)	.6755 ± .0250(4)	.7231 ± .0239(2)
HAB(80:20)	.7452 ± .0508(3)	.7695 ± .0499(1)	.7418 ± .0500(4)	.7627 ± .0486(2)
HAM(80:20)	.6587 ± .1110(3)	.7579 ± .0771(1)	.5640 ± .1061(4)	.7279 ± .0786(2)
LEV(80:20)	.8571 ± .0220(3)	.8766 ± .0197(1)	.8343 ± .0205(4)	.8630 ± .0183(2)
MMG(80:20)	.8090 ± .0243(3)	.8250 ± .0222(1)	.8051 ± .0236(4)	.8186 ± .0223(2)
MPG(80:20)	.9162 ± .0288(3)	.9326 ± .0219(1)	.9095 ± .0295(4)	.9235 ± .0260(2)
TAE(80:20)	.6950 ± .0814(3)	.7597 ± .0611(1)	.6663 ± .0788(4)	.7527 ± .0683(2)
YST(80:20)	.6905 ± .0221(3)	.7361 ± .0195(1)	.6800 ± .0230(4)	.7215 ± .0215(2)
Average ranking	3.00	1.00	4.00	2.00

Table 5: Mean and standard deviation of $\mathrm{Acc}_{U_{80}}$ (associated ranking) for CR, LR and their reliable variants, trained on 20%, 50%, and 80% of the datasets.

a subtler way makes it a useful variant of CR when cautious decision-making matters; ECR carries thus to CR the advantage of ELR over LR. In addition, the ability of ECR to capture nonlinear dependencies between input attributes makes it a useful alternative to ELR when accurate decision-making matters; ECR carries thus to ELR the advantage of CR over LR. Moreover, ECR compares also favorably with ELR in terms of reliable predictions, albeit to a lesser extent than in comparison to CR since the observed quantitative improvements were not confirmed by the statistical tests conducted. As will be seen, the experiment reported in the next section provides further evidence in favour of ECR over ELR in another setting related to cautious decision-making. Prior to that, we should nonetheless recall that these advantages come at a price: ECR combines also the drawbacks of CR and ELR in terms of computational complexity, as discussed in Section 4.3. Computing a prediction for ECR can indeed be intensive; this is typically the case for optimal (cross-validated) values of k such that $k \geq 8$.

5.4 Accuracy-rejection curves

In addition to the possibility of making imprecise predictions when there is too much uncertainty, ECR makes it possible to quantify the ignorance in a given prediction using the quantity $m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\mathcal{Y})$. In order to validate the relevance of this quantity as a measure of ignorance, we used accuracy-rejection curves [33], similarly as done previously in the context of reliable classification in [43]. This means that we rejected classification of instances for which $m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\mathcal{Y})$ is above some threshold, and evaluated the accuracy on the remaining instances. Then, we repeated this process for other thresholds. Finally, we plotted the obtained results as an accuracy-rejection curve, which is nothing but the accuracy on the non-rejected samples as a function of the rejection rate. The rationale for computing such curves is that, if $m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\mathcal{Y})$ is a valid measure of the ignorance associated with the classification of an object $\boldsymbol{x}$, then we should observe a monotone dependency between rejection rate and accuracy.

We repeated the experimental procedure described in Section 5.2, except that instead of looking at the misclassification rate on a given test set, we computed on this set the accuracy-rejection curve based on $m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\mathcal{Y})$ issued by ECR. Furthermore, we only considered the proportion 50:50 between training and test data. For completeness, we also computed accuracy-rejection curves based on $m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\mathcal{Y})$ issued by ELR, and we plotted accuracy-rejection curves based on random rejection (instead of $m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\mathcal{Y})$) of test instances classified by both ECR and ELR.

The results of this third experiment are shown in Figure 7. As it can be seen, for ECR, the accuracy on non-rejected instances is increasing in the rejection rate. This is an empirical confirmation that ECR is self-aware, through the quantity $m^{\mathcal{Y}}_{\mathcal{D},\boldsymbol{x}}(\mathcal{Y})$, of the reliability of its predictions. A similar observation can be made for ELR, although there are two noticeable differences. First, on the dataset TAE (see Fig. 7h), there is a drop in accuracy for ELR after the 50% rejection mark. This can perhaps be explained by the fact that this mark on this dataset corresponds to only 38 non-rejected instances, and thus the results past this mark may not be too meaningful. Second, on the dataset HAM (see Fig. 7d), ELR fail to do better than random rejection up to the 65% rejection mark. In contrast, ECR does not exhibit these undesirable behaviours. Moreover, its accuracy for any rejection rate is on average at least as good, or even better, than that of ELR. In sum, the flexibility of ECR with respect to ELR is also beneficial in terms of accuracy-rejection curves.

6 Conclusions

Choquistic regression is a generalisation of logistic regression, allowing flexibility in the modelling of interactions between predictors, while maintaining monotonicity. In this paper, we derived a reliable variant of choquistic regression, which makes it possible to represent the ignorance involved in a given prediction. This ability has been recognised as particularly important in critical classification problems such as medical diagnosis [43]. Our approach follows the one in [31], where a reliable version of logistic regression is proposed. It is formalised within evidence theory and relies in particular on a sound method for statistical inference and prediction developed in this framework. Through a series of experiments we showed that, by capturing uncertainty in a proper way and by enabling flexibility in the interactions between predictors, the reliable version of choquistic regression constitutes a useful alternative, when

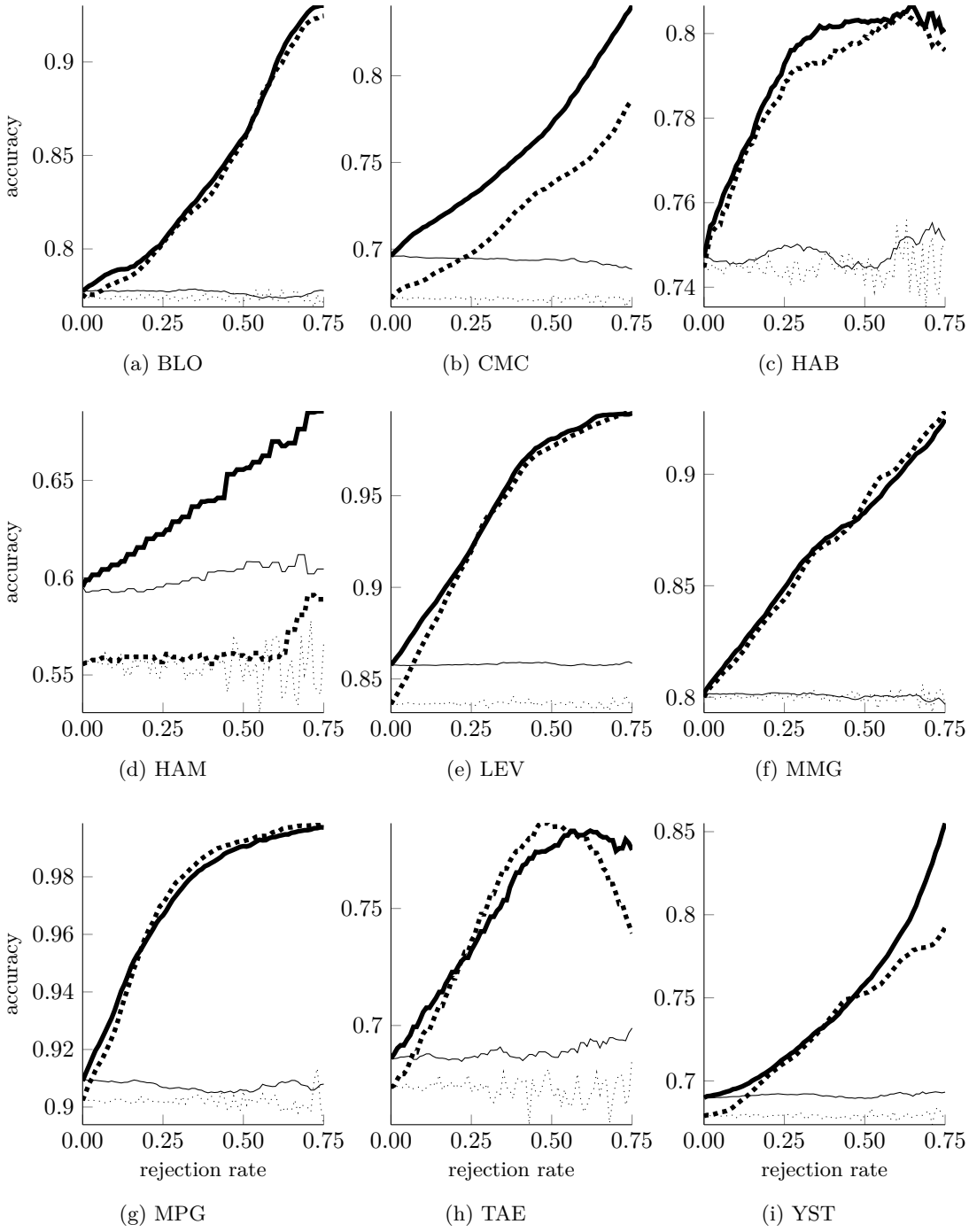


Figure 7: Accuracy-rejection curves based on the ignorance of ECR (▬) and of ELR (┅), and based on randomly rejecting test instances classified by ECR (–) and by ELR (···).

cautiousness in decision-making is allowed, both to choquistic regression and to the reliable logistic regression from [31].

The ability to quantify the ignorance, *aka* epistemic uncertainty, in a prediction seems particularly interesting in an active learning [44] setting, specifically in the context of the popular uncertainty sampling approach, as preliminary yet promising studies [41, 35] in this direction show. The rationale is that since ignorance corresponds to the reducible part of the uncertainty, it seems reasonable to assume that the active learner should query the label of the instance for which its current prediction has maximum ignorance, as this should lead to the most effective reduction of the learner's uncertainty, which is the typical goal of active learning. This was coined as the "epistemic uncertainty principle" in [35]. Ongoing work consists in testing this principle with respect to our reliable variant of choquistic regression.

The reliable versions of logistic regression proposed by Senge *et al.* [43] and by Minary *et al.* [31] seem to have been developed independently from one another. In future works, we plan on comparing these two approaches, both theoretically and empirically. Moreover, a belief function-based analysis of logistic regression was proposed recently in [14]. From a formal point of view, this analysis makes it possible to derive from the logistic regression model, a belief function on the unknown label of an object, and bears thus some similarity with the approach of Minary *et al.* [31]. We are interested by conducting also a thorough comparison between these two approaches. Such comparisons are not only interesting in themselves but could also inform further research in the direction of this paper, that is, developing reliable variants of generalisations of logistic regression.

Other perspectives include investigating two generalisations of our proposal. First, we would like to go beyond binary classification and handle multi-class classification. This may be achieved by following two different paths: (1) using binary decomposition, similarly as done in [36] to extend the reliable (binary) classification approach of Senge *et al.* [43], and in which case we could rely on previous works dealing with binary decomposition in the belief function framework [40, 27]; (2) by defining a multinomial version of choquistic regression, in a similar manner as multinomial logistic regression is defined (see, *e.g.*, [14]), and by deriving its belief function-based reliable variant using previous results for the multinomial logistic regression [48]. A second worthy generalisation would be the ability to handle learning data having soft labels (*aka* partially supervised learning). The foundations for such endeavour could be found in promising results concerning logistic regression obtained recently in [39].

Acknowledgements

This work is funded in part by the ELSAT2020 project, which is co-financed by the European Union with the European Regional Development Fund, the French state and the Hauts de France Region Council.

References

[1] N. Barile. *Studies in Learning Monotonic Models from Data.* PhD thesis, Utrecht University, Netherlands, 2014.

[2] P. H. Calamai and J. J. Moré. Projected gradient methods for linearly constrained problems. *Mathematical Programming*, 39(1):93–116, 1987.

[3] J.-R. Cano, N. R. Aljohani, R. A. Abbasi, J. S. Alowidbi, and S. García. Prototype selection to improve monotonic nearest neighbor. *Engineering Applications of Artificial Intelligence*, 60:128 – 135, 2017.

[4] J.-R. Cano, P. A. Gutiérrez, B. Krawczyk, M. Woźniak, and S. García. Monotonic classification: An overview on algorithms, performance measures and data sets. *Neurocomputing*, 341:168 – 182, 2019.

[5] N. Chakpitak, W. Yamaka, and S. Sriboonchitta. Comparing linear and nonlinear models in forecasting telephone subscriptions using likelihood based belief functions. In V. Kreinovich, S. Sriboonchitta, and N. Chakpitak, editors, *Predictive Econometrics and Big Data*, pages 363–374. Springer International Publishing, 2018.

[6] G. Choquet. Theory of capacities. *Annales de l'institut Fourier*, 5:131–295, 1954.

[7] G. Corani and M. Zaffalon. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *Journal of Machine Learning Research*, 9:581–621, June 2008.

[8] A. P. Dempster. New methods for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics*, 37(2):355–374, 1966.

[9] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[10] T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7):1095–1107, 1997.

[11] T. Denœux. Likelihood-based belief function: justification and some extensions to low-quality data. *International Journal of Approximate Reasoning*, 55(7):1535–1547, 2014.

[12] T. Denœux. Quantifying predictive uncertainty using belief functions: different approaches and practical construction. In V. Kreinovich, S. Sriboonchitta, and N. Chakpitak, editors, *Predictive Econometrics and Big Data*, pages 157–176. Springer International Publishing, 2018.

[13] T. Denœux. Decision-making with belief functions: a review. *International Journal of Approximate Reasoning*, 109:87–110, 2019.

[14] T. Denœux. Logistic regression, neural networks and Dempster-Shafer theory: a new perspective. *Knowledge-Based Systems*, 176:54–67, 2019.

[15] T. Denœux and O. Kanjanatarakul. Multistep prediction using point-cloud approximation of continuous belief functions. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6, June 2019.

[16] R. Duda and P. Hart. *Pattern recognition and scene analysis*. Wiley, New York, 1973.

[17] A. Fallah Tehrani, W. Cheng, K. Dembczyński, and E. Hüllermeier. Learning monotone nonlinear models using the Choquet integral. *Machine Learning*, 89(1):183–211, 2012.

[18] A. Feelders. Monotone relabeling in ordinal classification. In G. Webb, B. Liu, C. Zhang, D. Gunopulos, and X. Wu, editors, *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM)*, pages 803–808. IEEE Computer Society, 2010.

[19] J. Fodor and M. Roubens. *Fuzzy Preference Modelling and Multicriteria Decision Support*. Theory and Decision Library. Springer Netherlands, 1994.

[20] I. Gilboa and D. Schmeidler. Additive representations of non-additive measures and the Choquet integral. *Annals of Operations Research*, 52(1):43–65, 1994.

[21] M. Grabisch and C. Labreuche. A decade of application of the Choquet and Sugeno integrals in multi-criteria decision aid. *Annals of Operations Research*, 175:247–286, 2008.

[22] D. W. Hosmer and S. Lemeshow. *Applied logistic regression*. John Wiley & Sons, Ltd, 2000.

[23] J. Jaccard. *Interaction Effects in Logistic Regression*. Quantitative Applications in the Social Sciences. SAGE Publications, 2001.

[24] O. Kanjanatarakul, T. Denœux, and S. Sriboonchitta. Prediction of future observations using belief functions: A likelihood-based approach. *International Journal of Approximate Reasoning*, 72:71–94, 2016.

[25] O. Kanjanatarakul, S. Sriboonchitta, and T. Denœux. Forecasting using belief functions: An application to marketing econometrics. *International Journal of Approximate Reasoning*, 55(5):1113–1128, 2014.

[26] W. Kotłowski and R. Słowiundefinedski. Rule learning with monotonicity constraints. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML'09, page 537–544, New York, NY, USA, 2009. Association for Computing Machinery.

[27] M. Lachaize, S. L. Hégarat-Mascle, E. Aldea, A. Maitrot, and R. Reynaud. Evidential framework for error correcting output code classification. *Engineering Applications of Artificial Intelligence*, 73:10–21, 2018.

[28] L. Ma and T. Denœux. Making set-valued predictions in evidential classification: A comparison of different approaches. In J. De Bock, C. P. de Campos, G. de Cooman, E. Quaeghebeur, and G. Wheeler, editors, *Proceedings of the Eleventh International Symposium on Imprecise Probabilities: Theories and Applications*, volume 103 of *Proceedings of Machine Learning Research*, pages 276–285, 2019.

[29] N. Min, J. Sirisrisakulchai, and S. Sriboonchitta. Forecasting tourist arrivals to thailand using belief functions. In V.-N. Huynh, V. Kreinovich, S. Sriboonchitta, and K. Suriya, editors, *Econometrics of Risk*, pages 343–357. Springer International Publishing, 2015.

[30] P. Minary, F. Pichon, D. Mercier, E. Lefèvre, and B. Droit. Face pixel detection using evidential calibration and fusion. *International Journal of Approximate Reasoning*, 91:202–215, 2017.

[31] P. Minary, F. Pichon, D. Mercier, E. Lefèvre, and B. Droit. Evidential joint calibration of binary SVM classifiers. *Soft Computing*, 23(13):4655–4671, 2019.

[32] T. Minka. A comparison of numerical optimizers for logistic regression. Technical report, Carnegie Mellon University, Department of Statistics, 2003.

[33] M. Nadeem, J.-D. Zucker, and B. Hanczar. Accuracy-rejection curves (ARCs) for comparing classification methods with a reject option. *Journal of Machine Learning Research - Proceedings Track*, 8:65–81, 2010.

[34] H. T. Nguyen. On random sets and belief functions. *Journal of Mathematical Analysis and Applications*, 65(3):531–542, 1978.

[35] V.-L. Nguyen, S. Destercke, and E. Hüllermeier. Epistemic uncertainty sampling. In P. K. Novak, T. Smuc, and S. Dzeroski, editors, *Discovery Science - 22nd International Conference, DS 2019, Split, Croatia, October 28-30, 2019, Proceedings*, volume 11828 of *Lecture Notes in Computer Science*, pages 72–86. Springer, 2019.

[36] V.-L. Nguyen, S. Destercke, M.-H. Masson, and E. Hüllermeier. Reliable multi-class classification based on pairwise epistemic and aleatoric uncertainty. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 5089–5095. AAAI Press, 2018.

[37] P. Phochanachan, J. Sirisrisakulchai, and S. Sriboonchitta. Estimating oil price value at risk using belief functions. In V.-N. Huynh, V. Kreinovich, S. Sriboonchitta, and K. Suriya, editors, *Econometrics of Risk*, pages 377–389. Springer International Publishing, 2015.

[38] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large-Margin Classifiers*, 10:61–74, 1999.

[39] B. Quost, T. Denœux, and S. Li. Parametric classification with soft labels using the evidential EM algorithm: linear discriminant analysis versus logistic regression. *Advances in Data Analysis and Classification*, 11(4):659–690, 2017.

[40] B. Quost, T. Denœux, and M. Masson. Pairwise classifier combination using belief functions. *Pattern Recognition Letters*, 28(5):644–653, 2007.

[41] S. Ramel, F. Pichon, and F. Delmotte. Active evidential calibration of binary SVM classifiers. In S. Destercke, T. Denœux, F. Cuzzolin, and A. Martin, editors, *Belief Functions: Theory and Applications - 5th International Conference, BELIEF 2018, Compiègne, France, September 17-21, 2018, Proceedings*, volume 11069 of *Lecture Notes in Computer Science*, pages 208–216. Springer International Publishing, 2018.

[42] Y. U. Ryu, R. Chandrasekaran, and V. S. Jacob. Breast cancer prediction using the isotonic separation technique. *European Journal of Operational Research*, 181(2):842 – 854, 2007.

[43] R. Senge, S. Bösner, K. Dembczyński, J. Haasenritter, O. Hirsch, N. Donner-Banzhoff, and E. Hüllermeier. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29, 2014.

[44] B. Settles. *Active learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012.

[45] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, 1976.

[46] J. Sill. Monotonic networks. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 661–667. MIT Press, 1998.

[47] J. Sill and Y. S. Abu-Mostafa. Monotonicity hints. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, NIPS'96, page 634–640, Cambridge, MA, USA, 1996. MIT Press.

[48] P. Xu, F. Davoine, and T. Denœux. Evidential multinomial logistic regression for multiclass classifier calibration. In *18th International Conference on Information Fusion, FUSION 2015, Washington, DC, USA, July 6-9, 2015*, pages 1106–1112. IEEE, 2015.

[49] P. Xu, F. Davoine, H. Zha, and T. Denœux. Evidential calibration of binary SVM classifiers. *International Journal of Approximate Reasoning*, 72:55–70, 2016.

[50] G. Yang, S. Destercke, and M.-H. Masson. The costs of indeterminacy: How to determine them? *IEEE Transactions on Cybernetics*, 47(12):4316–4327, 2017.

[51] M. Zaffalon, G. Corani, and D. Mauá. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282–1301, 2012.

[52] J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, 14(1):185–205, 2005.