Knowledge-Based Systems 176 (2019) 54–67

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Logistic regression, neural networks and Dempster–Shafer theory: A new perspective

Thierry Denœux

Université de Technologie de Compiègne, CNRS UMR 7253 Heudiasyc, Compiègne, France

ARTICLE INFO

Article history:
Received 8 February 2019
Received in revised form 23 March 2019
Accepted 25 March 2019
Available online 27 March 2019

Keywords:
Classification
Pattern recognition
Supervised learning
Evidence theory
Belief functions

ABSTRACT

We revisit logistic regression and its nonlinear extensions, including multilayer feedforward neural networks, by showing that these classifiers can be viewed as converting input or higher-level features into Dempster–Shafer mass functions and aggregating them by Dempster's rule of combination. The probabilistic outputs of these classifiers are the normalized plausibilities corresponding to the underlying combined mass function. This mass function is more informative than the output probability distribution. In particular, it makes it possible to distinguish between lack of evidence (when none of the features provides discriminant information) from conflicting evidence (when different features support different classes). This expressivity of mass functions allows us to gain insight into the role played by each input feature in logistic regression, and to interpret hidden unit outputs in multilayer neural networks. It also makes it possible to use alternative decision rules, such as interval dominance, which select a set of classes when the available evidence does not unambiguously point to a single class, thus trading reduced error rate for higher imprecision.

1. Introduction

The Dempster–Shafer (DS) theory of belief functions [1,2] is now well-established as a formalism for reasoning and making decisions with uncertainty [3]. DS theory, also referred to as *Evidence Theory*, is essentially based on representing independent pieces of evidence by completely monotone capacities (also called belief functions), and pooling them using a generic operator called Dempster's rule of combination.

In the last twenty years, DS theory has been increasingly applied to statistical pattern recognition and, in particular, to supervised classification. One direction of research is *classifier fusion*: classifier outputs are expressed as belief functions and combined by Dempster's rule or any other rule (see, e.g., [4–9]). Another approach is *evidential calibration*, which converts the decisions of statistical classifiers (such as support vector machines) into belief functions [10–12]. The third approach, which is maybe the most promising and the focus of this paper, is to design *evidential classifiers*, whose basic principles are rooted in DS theory. Typically, an evidential classifier breaks down the evidence of each input feature vector into elementary mass functions and combines them by Dempster's rule. The combined mass function (or *orthogonal sum*) can then be used for decision-making [13]. Thanks to the generality and expressiveness of the belief function formalism, evidential classifiers provide more informative outputs than those of conventional classifiers. This expressiveness can be exploited, in particular, for uncertainty quantification, novelty detection and information fusion in decision-aid or fully automatic decision systems.

Over the years, several principles for designing evidential classifiers have been developed. In [14], a distinction was made between the so-called *model-based* approach, which uses estimated class-conditional distributions and the "Generalized Bayes Theorem", an extension of Bayes theorem [15,16], and the *case-based*, or *distance-based* approach, in which mass functions m_j are constructed based on distances to learning instances or to prototypes. Evidential classifiers in the latter category have been used in a wide range of applications [17–19]. They include the *evidential k-nearest neighbor rule* [20] and its variants (see, e.g. [21–25]), as well as the *evidential neural network classifier* [26], in which mass functions are constructed based on the distances to prototypes, and the whole system is trained to minimize an error function.

In this paper, we show that not only these particular model-based and distance-based classifiers, but also a broad class of supervised machine learning algorithms, can be seen as evidential classifiers. This class contains logistic regression and its nonlinear generalizations, including multilayer feedforward neural networks, generalized additive models, support vector machines and, more generally, all classifiers based on linear combinations of input or higher-order features and their transformation through the logistic or softmax transfer function. We will show that *generalized logistic regression classifiers* can be seen as computing the orthogonal sum of elementary pieces of evidence supporting each

E-mail address: thierry.denoeux@utc.fr.

https://doi.org/10.1016/j.knosys.2019.03.030
0950-7051/

class or its complement. The output class probabilities are then normalized plausibilities according to some underlying DS mass function, the expression of which is laid bare in this paper. This "hidden" mass function provides a more informative description of the classifier output than the class probabilities, and can be used for decision-making. Also, the individual mass functions computed by each of the features provides insight into the internal operation of classifier and can help to interpret its decisions. This finding leads us to the conclusion that DS theory is a much more general framework for classifier analysis and construction than was initially believed, and opens a new perspective for the study and practical application of a wide range of machine learning algorithms.[1]

The rest of this paper is organized as follows. DS theory and some principles of classifier construction will first be recalled in Section 2. The new connection between DS theory and some machine learning models will then be established in Section 3, and the identification of DS model will be addressed in Section 4. Finally, some numerical experiments will be presented in Section 5, and Section 6 will conclude the paper.

2. Background

In this section, we first recall some necessary definitions and results from DS theory (Section 2.1). We then provide brief descriptions of logistic regression and neural network classifiers that will be considered later in the paper (Section 2.2).

2.1. Dempster–Shafer theory

2.1.1. Mass function

Let $\Theta = \{\theta_1, \ldots, \theta_K\}$ be a finite set. A *mass function* on Θ is a mapping $m : 2^\Theta \to [0, 1]$ such that $m(\emptyset) = 0$ and

$$\sum_{A \subseteq \Theta} m(A) = 1.$$

In DS theory, Θ is the set of possible answers to some question, and a mass function m represent a piece of evidence pertaining to that question. Each mass $m(A)$ represents a share of a unit mass of belief allocated to the hypothesis that the truth is in A, and which cannot be allocated to any strict subset of A. Each subset $A \subseteq \Theta$ such that $m(A) > 0$ is called a *focal set* of m. A mass function m is said to be *simple* if it has the following form:

$$m(A) = s, \quad m(\Theta) = 1 - s, \tag{1}$$

for some $A \subset \Theta$ such that $A \neq \emptyset$ and some $s \in [0, 1]$, called the *degree of support* in A. For a reason that will become apparent later, the quantity $w := -\ln(1 - s)$ is called the *weight of evidence*[2] associated to m [2, page 77]. The *vacuous* mass function, corresponding to $s = w = 0$, represents complete ignorance.

2.1.2. Belief and plausibility functions

Given a mass function m, *belief* and *plausibility* functions are defined, respectively, as follows:

$$Bel(A) := \sum_{B \subseteq A} m(B) \tag{2a}$$

$$Pl(A) := \sum_{B \cap A \neq \emptyset} m(B) = 1 - Bel(\bar{A}), \tag{2b}$$

for all $A \subseteq \Theta$. The quantity $Bel(A)$ can be interpreted as the degree of total support to A, while $1 - Pl(A)$ is the degree of total support to $\bar{A}$, i.e., the degree of doubt in A [2]. The *contour function* $pl : \Theta \to [0, 1]$ is the restriction of the plausibility function Pl to singletons, i.e., $pl(\theta) = Pl(\{\theta\})$, for all $\theta \in \Theta$.

2.1.3. Dempster's rule

Two mass functions m_1 and m_2 representing independent items of evidence can be combined using Dempster's rule [1,2] defined as

$$(m_1 \oplus m_2)(A) := \frac{1}{1 - \kappa} \sum_{B \cap C = A} m_1(B) m_2(C), \tag{3}$$

for all $A \subseteq \Theta$, $A \neq \emptyset$, and $(m_1 \oplus m_2)(\emptyset) := 0$. In (3), κ is the *degree of conflict* between the two mass functions, defined as

$$\kappa := \sum_{B \cap C = \emptyset} m_1(B) m_2(C). \tag{4}$$

Mass function $m_1 \oplus m_2$ is well defined if $\kappa < 1$. It is then called the *orthogonal sum* of m_1 and m_2. Dempster's rule is commutative and associative, and the vacuous mass function is its only neutral element. The contour function $pl_1 \oplus pl_2$ associated to $m_1 \oplus m_2$ can be computed as

$$pl_1 \oplus pl_2(\theta) = \frac{pl_1(\theta) pl_2(\theta)}{1 - \kappa}, \tag{5}$$

for all $\theta \in \Theta$.

2.1.4. Weights of evidence

Given two simple mass functions m_1 and m_2 with the same focal set A and degrees of support s_1 and s_2, their orthogonal sum is the simple mass function

$$(m_1 \oplus m_2)(A) = 1 - (1 - s_1)(1 - s_2) \tag{6a}$$

$$(m_1 \oplus m_2)(\Theta) = (1 - s_1)(1 - s_2). \tag{6b}$$

The corresponding weight of evidence is, thus,

$$w = -\ln[(1 - s_1)(1 - s_2)] \tag{7a}$$

$$= -\ln(1 - s_1) - \ln(1 - s_2) = w_1 + w_2, \tag{7b}$$

i.e., weights of evidence add up when aggregating evidence using Dempster's rule. Denoting a simple mass function with focal set A and weight of evidence w as A^w, this property can be expressed by the following equation,

$$A^{w_1} \oplus A^{w_2} = A^{w_1 + w_2}. \tag{8}$$

We note that, in [29], following [30], we used the term "weight" for $-\ln w$. As we will see, the additivity property is central in our analysis: we thus stick to Shafer's terminology and notation in this paper. A mass function is said to be *separable* if it can be decomposed as the orthogonal sum of simple mass functions [2, page 87]. A separable mass function can thus be written as

$$m = \bigoplus_{\emptyset \neq A \subset \Theta} A^{w(A)},$$

where $w(\cdot)$ is a mapping from $2^\Theta \setminus \{\emptyset, \Theta\}$ to $[0, +\infty)$.

2.1.5. Plausibility transformation

It is sometimes useful to approximate a DS mass function m by a probability mass function $p_m : \Theta \to [0, 1]$. One such approximation with good properties is obtained by normalizing the contour function [31,32]; we then have

$$p_m(\theta_k) := \frac{pl(\theta_k)}{\sum_{l=1}^K pl(\theta_l)}, \quad k = 1, \ldots, K. \tag{9}$$

[1] A preliminary version of this paper with some partial results appeared as a short conference paper [27].

[2] This notion of "weight of evidence" in DS theory should not be confused with related, but different notions with similar names proposed in other contexts such as rough set theory, as reviewed in [28].

As a consequence of (5), the so-called *plausibility transformation* (9) has the following interesting property in relation with Dempster's rule:

$$p_{m_1 \oplus m_2}(\theta_k) \propto p_{m_1}(\theta_k) p_{m_2}(\theta_k), \quad k = 1, \ldots, K,$$

i.e., the probability distribution associated to $m_1 \oplus m_2$ can be computed in $O(K)$ arithmetic operations by multiplying the probability distributions p_{m_1} and p_{m_2} elementwise, and renormalizing.

2.1.6. Least commitment principle

The *maximum uncertainty* [33] or *least commitment* [15] principle serves the same purpose as the maximum entropy principle in probability theory. According to this principle, when several belief functions are compatible with a set of constraints, the least committed (or informative) should be selected. In order to apply this principle, we need to define a partial order on the set of belief functions. For that purpose, we may either define a degree of imprecision or uncertainty of a belief function [33], or we may adopt a more qualitative approach and directly define an informational ordering relation on the set of belief functions [34,35].

If we restrict ourselves to separable mass functions, as will be done in this paper, we can compare mass functions by their weights of evidence. Given two separable mass functions $m_1 = \bigoplus_{\emptyset \neq A \subset \Theta} A^{w_1(A)}$ and $m_2 = \bigoplus_{\emptyset \neq A \subset \Theta} A^{w_2(A)}$, it makes sense to consider that m_1 is more committed than m_2 (denoted as $m_1 \sqsubseteq_w m_2$) if it has larger weights of evidence, i.e, if $w_1(A) \geq w_2(A)$ for all A [29]. Because of (8), combining m_2 with a separable mass function m results in a more committed mass function $m_1 = m_2 \oplus m$, with $m_1 \sqsubseteq_w m_2$.

A related family of measures of information content is defined by

$$I_p(m) := \sum_{\emptyset \neq A \subset \Theta} w(A)^p, \quad p > 0. \tag{10}$$

Clearly, for any two separable mass functions m_1 and m_2, $m_1 \sqsubseteq_w m_2 \Rightarrow I_p(m_1) \geq I_p(m_2)$.

2.1.7. Decision analysis

Consider a decision problem with a set $\mathcal{A} = \{a_1, \ldots, a_r\}$ of acts, a set $\Theta = \{\theta_1, \ldots, \theta_K\}$ of states of nature, and a loss function $L : \mathcal{A} \times \Theta \to \mathbb{R}$. The lower and upper risks of act a with respect to a mass function m are defined, respectively, as the lower and upper expected loss [1,36], if the decision-maker (DM) selects act a:

$$R_*(a) := \sum_{A \subseteq \Theta} m(A) \min_{\theta \in A} L(a, \theta),$$

$$R^*(a) := \sum_{A \subseteq \Theta} m(A) \max_{\theta \in A} L(a, \theta).$$

A pessimistic (resp., optimistic) DM will prefer act a over a' if $R^*(a) \leq R^*(a')$ (resp., $R_*(a) \leq R_*(a')$). Alternatively, a conservative approach is to consider a preferable to a' whenever $R^*(a) \leq R_*(a')$. This *interval dominance (ID)* preference relation [37] is a partial preorder on $\mathcal{A}$. For decision-making, one can select the set of maximal elements of this relation, defined as $\{a \in \mathcal{A} | \forall a' \in \mathcal{A} \setminus \{a\}, R^*(a') > R_*(a)\}$. In classification, act a_k is usually interpreted as selecting class k, and we have $r = K$. Assuming the 0–1 loss function defined by $L(a_k, \theta_l) = 1 - \delta_{kl}$, where δ is the Kronecker delta, we have $R_*(a_k) = 1 - pl(\theta_k)$ and $R^*(a_k) = 1 - Bel(\{\theta_k\})$. The optimistic rule then selects the class with the highest plausibility [13]. This rule will be hereafter referred to as the *maximum plausibility (MP)* rule.

2.2. Logistic regression

In the following, we recall some basic definitions and notations about classification. We start with binary logistic regression and proceed with the multi-category case and some nonlinear extensions.

2.2.1. Binary logistic regression

Consider a binary classification problem with d-dimensional feature vector $X = (X_1, \ldots, X_d)$ and class variable $Y \in \Theta = \{\theta_1, \theta_2\}$. Let $p_1(x)$ denote the probability that $Y = \theta_1$ given that $X = x$. In the binary logistic regression model, it is assumed that

$$\ln \frac{p_1(x)}{1 - p_1(x)} = \beta^T x + \beta_0, \tag{11}$$

where $\beta \in \mathbb{R}^d$ and $\beta_0 \in \mathbb{R}$ are parameters. Solving (11) for $p_1(x)$, we get

$$p_1(x) = \frac{1}{1 + \exp[-(\beta^T x + \beta_0)]}. \tag{12}$$

Given a learning set $\{(x_i, y_i)\}_{i=1}^n$, parameters β and β_0 are usually estimated by maximizing the conditional log-likelihood

$$\ell(\beta, \beta_0) = \sum_{i=1}^{n} y_{i1} \ln p_1(x_i) + (1 - y_{i1}) \ln [1 - p_1(x_i)], \tag{13}$$

where $y_{i1} = 1$ if $y_i = \theta_1$ and $y_{i1} = 0$ otherwise.

2.2.2. Multinomial logistic regression

Consider now a multiclass classification problem with $K > 2$ classes, and let $\Theta = \{\theta_1, \ldots, \theta_K\}$ denote the set of classes. Multinomial logistic regression extends binary logistic regression by assuming the log-posterior probabilities to be affine functions of x:

$$\ln p_k(x) = \beta_k^T x + \beta_{k0} + \gamma, \quad k = 1, \ldots, K, \tag{14}$$

where $p_k(x) = \mathbb{P}(Y = \theta_k | X = x)$ is the posterior probability of class θ_k, $\beta_k \in \mathbb{R}^d$ and $\beta_{k0} \in \mathbb{R}$ are class-specific parameters and $\gamma \in \mathbb{R}$ is a constant that does not depend on k. The posterior probability of class θ_k can then be expressed as

$$p_k(x) = \frac{\exp(\beta_k^T x + \beta_{k0})}{\sum_{l=1}^{K} \exp(\beta_l^T x + \beta_{l0})}, \tag{15}$$

and parameters (β_k, β_{k0}), $k = 1 \ldots, K$ can be estimated by maximizing the conditional likelihood as in the binomial case. The transformation from linear combinations of features $\beta_k^T x + \beta_{k0} \in \mathbb{R}$ to probabilities in [0, 1] described by (15) is often referred to as the *softmax transformation*.

2.2.3. Nonlinear extensions

Logistic regression classifiers define decision regions separated by hyperplanes: they are linear classifiers. However, nonlinear classifiers can be built by applying logistic regression to transformed features $\phi_j(x)$, $j = 1, \ldots, J$, where the ϕ_j's are nonlinear mappings from $\mathbb{R}^d$ to $\mathbb{R}$. We call such classifiers *generalized logistic regression (GLR) classifiers* (see Fig. 1). Both the new features $\phi_j(x)$ and the coefficients (β_k, β_{k0}) are usually learnt simultaneously by minimizing some cost function. Popular models based on this principle include quadratic logistic regression [38], generalized additive models [39], multilayer feedforward neural networks [40,41], radial basis function networks [42] and support vector machines [43]. In particular, Feedforward Neural Networks (FNNs) are models composed of elementary computing units (or "neurons") arranged in layers. Each layer computes a vector of new features as functions of the outputs from the previous layer. For classification, the output layer is typically a softmax layer

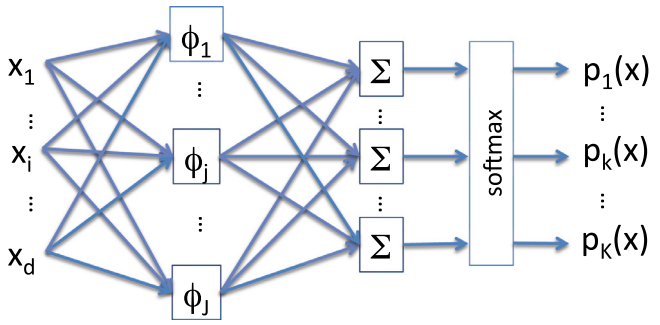


Fig. 1. Generalized logistic regression classifier.

with K output units. This model is thus equivalent to logistic regression performed on new features computed in the network's hidden layers. All weights in the network are learnt by minimizing a cost function, which is often taken as the negative conditional likelihood (or cross-entropy), as in logistic regression.

3. DS Analysis of GLR classifiers

In this section, we expose the main result of this paper, which establishes a bridge between DS theory, recalled in Section 2.1, and the GLR classifiers summarized in Section 2.2. We start with binary classification in Section 3.1, and proceed with the multi-category case in Section 3.2.

3.1. Case $K = 2$

Consider a binary classification problem with $K = 2$ classes in $\Theta = \{\theta_1, \theta_2\}$. Let $\phi(x) = (\phi_1(x), \ldots, \phi_J(x))$ be a vector of J features. These features may be the input features, in which case we have $\phi_j(x) = x_j$ for all j and $J = d$, or nonlinear functions thereof. Each feature value $\phi_j(x)$ is a piece of evidence about the class $Y \in \Theta$ of the instance under consideration. Assume that this evidence points either to θ_1 or θ_2, depending on the sign of

$$w_j := \beta_j \phi_j(x) + \alpha_j, \tag{16}$$

where β_j and α_j are two coefficients. The weights of evidence for θ_1 and θ_2 are assumed to be equal to, respectively, the positive part $w_j^+ := \max(0, w_j)$ of w_j, and its negative part $w_j^- := \max(0, -w_j)$. Under this model, the consideration of feature ϕ_j induces the simple mass function

$$m_j = \{\theta_1\}^{w_j^+} \oplus \{\theta_2\}^{w_j^-}.$$

3.1.1. Output mass function

Assuming that the values of the J features can be considered as independent pieces of evidence, the combined mass function after taking into account the J features is

$$m = \bigoplus_{j=1}^{J} \left(\{\theta_1\}^{w_j^+} \oplus \{\theta_2\}^{w_j^-} \right) \tag{17a}$$

$$= \left(\bigoplus_{j=1}^{J} \{\theta_1\}^{w_j^+} \right) \oplus \left(\bigoplus_{j=1}^{J} \{\theta_2\}^{w_j^-} \right) \tag{17b}$$

$$= \{\theta_1\}^{w^+} \oplus \{\theta_2\}^{w^-}, \tag{17c}$$

where $w^+ := \sum_{j=1}^{J} w_j^+$ and $w^- := \sum_{j=1}^{J} w_j^-$ are the total weights of evidence supporting, respectively, θ_1 and θ_2. Denoting by m^+ and m^- the two mass functions on the right-hand side of Eq. (17c), we have

$$m^+(\{\theta_1\}) = 1 - \exp(-w^+) \tag{18a}$$

$$m^+(\Theta) = \exp(-w^+) \tag{18b}$$

and

$$m^-(\{\theta_2\}) = 1 - \exp(-w^-) \tag{19a}$$

$$m^-(\Theta) = \exp(-w^-). \tag{19b}$$

Hence,

$$m(\{\theta_1\}) = \frac{[1 - \exp(-w^+)] \exp(-w^-)}{1 - \kappa} \tag{20a}$$

$$m(\{\theta_2\}) = \frac{[1 - \exp(-w^-)] \exp(-w^+)}{1 - \kappa} \tag{20b}$$

$$m(\Theta) = \frac{\exp(-w^+ - w^-)}{1 - \kappa} = \frac{\exp(-\sum_{j=1}^{J} |w_j|)}{1 - \kappa}, \tag{20c}$$

where

$$\kappa = [1 - \exp(-w^+)][1 - \exp(-w^-)] \tag{21}$$

is the degree of conflict between m^- and m^+. Mass function m defined Eqs (20) and (21) is the output of the evidential classifier. As shown in Fig. 2(a), $m(\{\theta_1\})$ is increasing w.r.t. w^+ and decreasing w.r.t. w^-, while the mass $m(\Theta)$ is a decreasing function of the total weight of evidence $w^+ + w^-$ (Fig. 2(b)). The degree of conflict κ increases with both w^- and w^+ (Fig. 2(c)).

3.1.2. Contour function

The contour function corresponding to m is

$$pl(\theta_1) = m(\{\theta_1\}) + m(\Theta) = \frac{\exp(-w^-)}{1 - \kappa} \tag{22a}$$

$$pl(\theta_2) = m(\{\theta_2\}) + m(\Theta) = \frac{\exp(-w^+)}{1 - \kappa}. \tag{22b}$$

We can observe that Eq. (22) is consistent with the semantics of plausibility: the plausibility of class θ_1 is high when there is little evidence in favor of θ_2, i.e., when w^- is low. Applying the plausibility transformation (9), we get the following probability of θ_1:

$$p_m(\theta_1) = \frac{\exp(-w^-)}{\exp(-w^-) + \exp(-w^+)} \tag{23a}$$

$$= \frac{1}{1 + \exp(w^- - w^+)} \tag{23b}$$

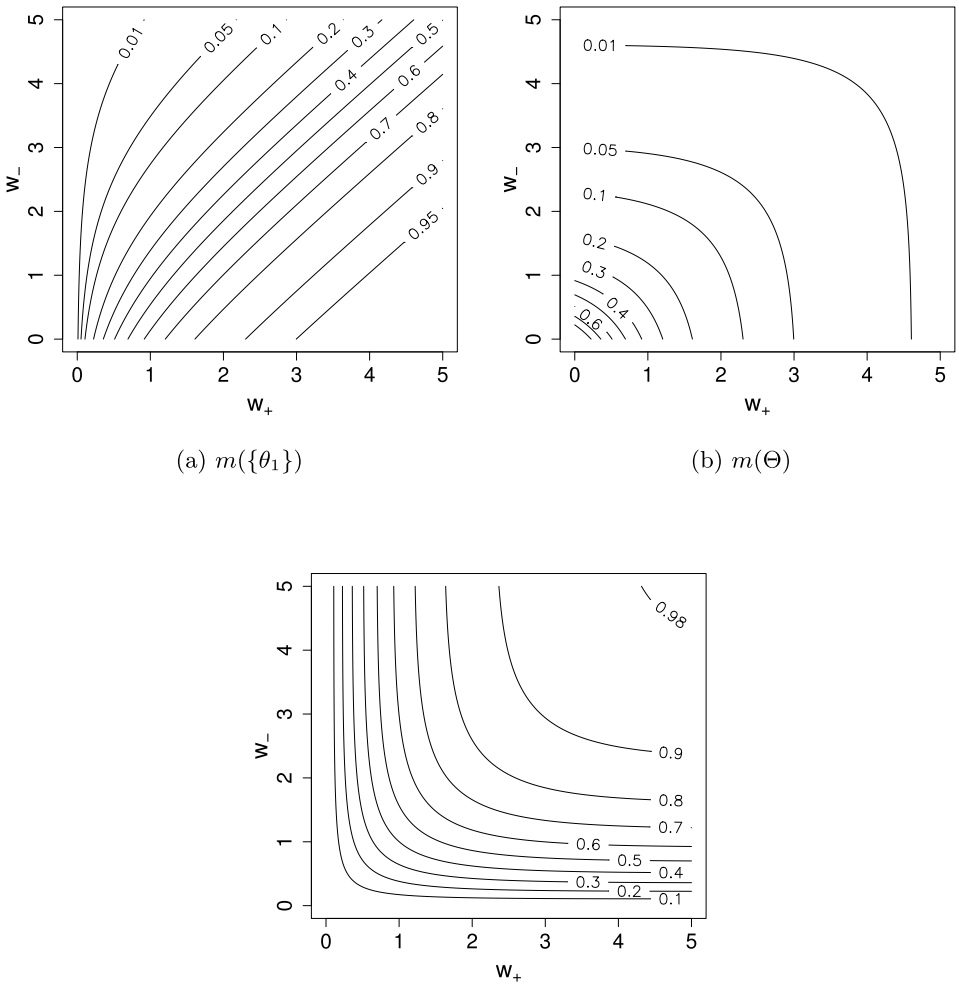
$$= \frac{1}{1 + \exp[-(\beta^T \phi(x) + \beta_0)]}, \tag{23c}$$

with $\beta = (\beta_1, \ldots, \beta_J)$ and

$$\beta_0 = \sum_{i=1}^{J} \alpha_j. \tag{24}$$

3.1.3. Discussion

We observe that (23c) is identical to (12): in the two-category case, the probabilities computed by logistic regression can, thus, be viewed as normalized plausibilities obtained by a process of evidence combination in the DS framework. Fig. 3 contrasts the classical view of binary logistic regression with the DS view outlined here. If one considers only the output probability, both views are equivalent. The latter, however, lays bare an underlying mass function m, which has one more degree of freedom than the output probability $p_1(x)$. This additional degree of freedom makes it possible to distinguish, e.g., between lack of evidence, in which



(a) $m(\{\theta_1\})$ (b) $m(\Theta)$

(c) Degree of conflict κ

Fig. 2. Contour lines of $m(\{\theta_1\})$ (a), $m(\Theta)$ (b) and the degree of conflict κ (c) as functions of w^+ (horizontal axis) and w^- (vertical axis).

case we have $m(\Theta) = 1$, and maximally conflicting evidence corresponding to $m(\{\theta_1\}) = m(\{\theta_2\}) = 0.5$. These two cases result in the same output probability $p_1(x) = 0.5$. This distinction has implications for decision making, as will be shown in Section 5.

Typically, parameters β and β_0 are estimated by maximizing the log-likelihood function (13), but parameters $\alpha_j, j = 1, \ldots, J$ are unidentifiable. We will return to this point in Section 4. Before that, we address the multi-category case in the following section.

3.2. Case $K > 2$

Let us now consider the multi-category case, where $K > 2$.

3.2.1. Model

For each θ_k, we now assume that the evidence of feature $\phi_j(x)$ points either to the singleton $\{\theta_k\}$ or to its complement $\overline{\{\theta_k\}}$, depending on the sign of

$$w_{jk} := \beta_{jk}\phi_j(x) + \alpha_{jk}, \tag{25}$$

where $(\beta_{jk}, \alpha_{jk})$, $k = 1, \ldots, K$, $j = 1, \ldots, J$ are parameters. The weights of evidence for $\{\theta_k\}$ and $\overline{\{\theta_k\}}$ are supposed to be equal, respectively, to the positive and negative parts of w_{jk}, denoted by w_{jk}^+ and w_{jk}^-, respectively. For each feature ϕ_j and each class θ_k, we thus have two simple mass functions, $m_{jk}^+ := \{\theta_k\}^{w_{jk}^+}$ and $m_{jk}^- := \overline{\{\theta_k\}}^{w_{jk}^-}$. Assuming these mass functions to be independent,

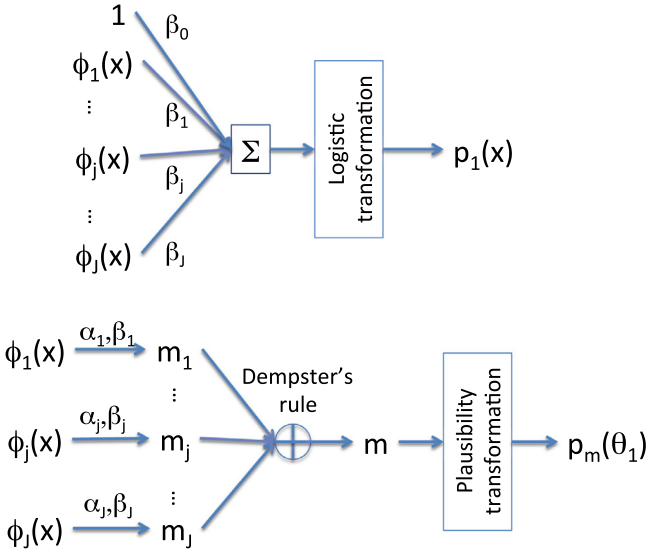


Fig. 3. Classical view (top) and DS view (bottom) of a binary GLR classifier.

they can be combined by Dempster's rule. Combining separately the positive and the negative evidence with respect to each class

θ_k, we get

$$m_k^+ := \bigoplus_{j=1}^{J} m_{jk}^+ = \{\theta_k\}^{w_k^+} \tag{26a}$$

$$m_k^- := \bigoplus_{j=1}^{J} m_{jk}^- = \overline{\{\theta_k\}}^{w_k^-}, \tag{26b}$$

where

$$w_k^+ := \sum_{j=1}^{J} w_{jk}^+ \quad \text{and} \quad w_k^- := \sum_{j=1}^{J} w_{jk}^-. \tag{27}$$

3.2.2. Combined contour function

The contour functions pl_k^+ and pl_k^- associated, respectively, with m_k^+ and m_k^- are

$$pl_k^+(\theta) = \begin{cases} 1 & \text{if } \theta = \theta_k, \\ \exp\left(-w_k^+\right) & \text{otherwise,} \end{cases}$$

and

$$pl_k^-(\theta) = \begin{cases} \exp\left(-w_k^-\right) & \text{if } \theta = \theta_k, \\ 1 & \text{otherwise.} \end{cases}$$

Now, let $m^+ = \bigoplus_{k=1}^{K} m_k^+$ and $m^- = \bigoplus_{k=1}^{K} m_k^-$ be the mass functions pooling, respectively, all the positive and the negative evidence, and let pl^+ and pl^- be the corresponding contour functions. From (5), we have

$$pl^+(\theta_k) \propto \prod_{l=1}^{K} pl_l^+(\theta_k) = \exp\left(-\sum_{l\neq k} w_l^+\right) = \exp\left(-\sum_{l=1}^{K} w_l^+\right)\exp(w_k^+) \propto \exp(w_k^+),$$

and

$$pl^-(\theta_k) \propto \prod_{l=1}^{K} pl_l^-(\theta_k) = \exp(-w_k^-). \tag{28}$$

Finally, let $m = m^+ \oplus m^-$ and let pl be the corresponding contour function. Using again Eq. (5), we have

$$pl(\theta_k) \propto pl^+(\theta_k)pl^-(\theta_k) \propto \exp(w_k^+ - w_k^-) \propto \exp\left(\sum_{j=1}^{J} w_{jk}\right) = \exp\left(\sum_{j=1}^{J} \beta_{jk}\phi_j(x) + \sum_{j=1}^{J} \alpha_{jk}\right).$$

Let p be the probability mass function induced from m by the plausibility transformation (9), and let

$$\beta_{0k} := \sum_{j=1}^{J} \alpha_{jk}. \tag{29}$$

The probability of class θ_k induced by mass function m is

$$p_m(\theta_k) = \frac{\exp\left(\sum_{j=1}^{J} \beta_{jk}\phi_j(x) + \beta_{0k}\right)}{\sum_{l=1}^{K} \exp\left(\sum_{j=1}^{J} \beta_{jl}\phi_j(x) + \beta_{0l}\right)}. \tag{30}$$

It is identical to (15). We thus have proved that the result found in Section 4.1 for the binary case also holds in the multi-category case: conditional class probabilities computed by a multinomial GLR classifier can be seen as the normalized plausibilities obtained after combining elementary mass functions $m_{jk} = m_{jk}^+ \oplus m_{jk}^-$ by Dempster's rule: these classifiers are, thus, evidential classifiers as defined in Section 1. The classical and DS views of multinomial GLR classifiers are contrasted in Fig. 4.

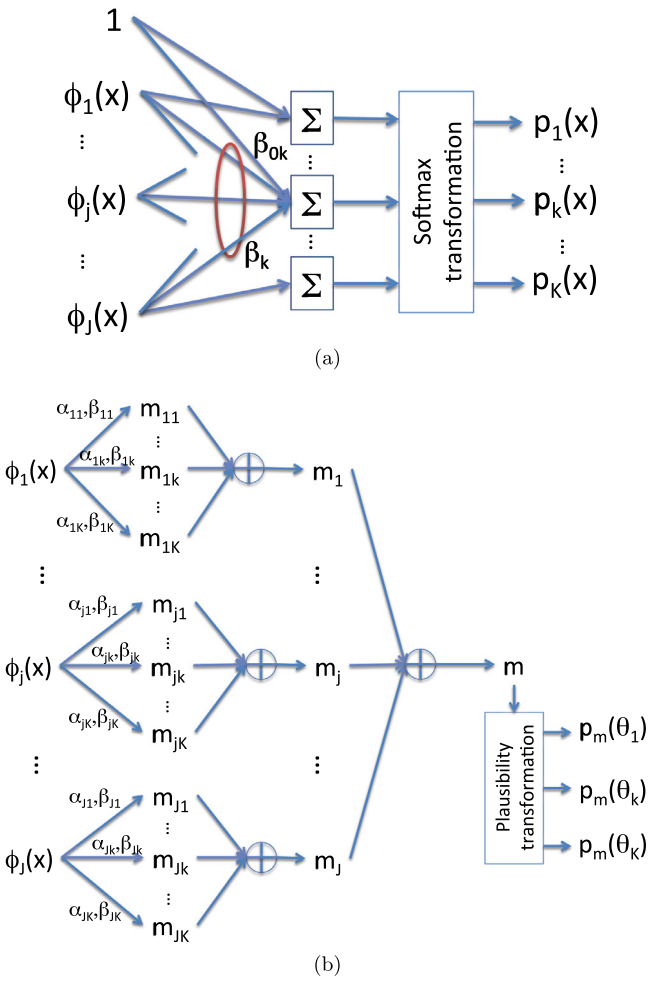


Fig. 4. Classical view (a) and DS view (b) of a multinomial GLR classifier.

3.2.3. Output mass function

As in the binary case, we can compute the expression of the underlying mass function m. Its expression in the multi-category case is more complex than it is in the binary case. It is given in the following proposition.

Proposition 1. *The output mass function*

$$m = \bigoplus_{k=1}^{K} \left(\{\theta_k\}^{w_k^+} \oplus \overline{\{\theta_k\}}^{w_k^-}\right) \tag{31}$$

has the following expression:

$$m(\{\theta_k\}) = C\exp(-w_k^-)\left\{\exp(w_k^+) - 1 + \prod_{l\neq k}\left[1 - \exp(-w_l^-)\right]\right\},$$

for $k = 1, \ldots, K$, *and*

$$m(A) = C\left\{\prod_{\theta_k \notin A}\left[1 - \exp(-w_k^-)\right]\right\}\left\{\prod_{\theta_k \in A}\exp(-w_k^-)\right\}$$

for any $A \subseteq \Theta$ *such that* $|A| > 1$, *where* C *is a proportionality constant.*

Proof. see Appendix A.

4. Identification of model parameters

To compute the output mass function given by Eq. (20) in the binary case and by Proposition 1 in the multi-category case, we need to compute the weights of evidence. In the binary case, these weights depend on coefficients β_j and α_j for $j = 1, \ldots, J$ through (16). A learning procedure (such as likelihood maximization) gives us estimates $\widehat{\beta}_j$ of β_j for $j = 0, \ldots, J$. Parameters α_j are not identifiable, but are linked to β_0 by Eq. (24). In the multi-category case, things are worse, because parameters β_{jk} are also not identifiable: we can easily check that adding any constant vector $\boldsymbol{c} = (c_0, \ldots, c_J)$ to each vector $\beta_k = (\beta_{0k}, \ldots, \beta_{Jk})$ produces the same normalized plausibilities (30). Both parameters β_{jk} and α_{jk} are, thus, underdetermined in that case.

To identify the model parameters, we propose to apply the Least Commitment Principle introduced in Section 2.1.6, by searching for the parameter values that give us the output mass functions with minimal information content, the information content of a mass function m being taken to be $I_p(m)$ defined by (10), with $p = 2$. (The value $p = 2$ is chosen because it lends itself to easy computation, as will be shown below). We will first deal with the binary case in Section 4.1 and proceed with the multi-category case in Section 4.2.

4.1. Binary case

Let $\{(x_i, y_i)\}_{i=1}^n$ be the learning set, let $\widehat{\beta}_j$ be the maximum likelihood estimate of the coefficients β_j, and let $\boldsymbol{\alpha}$ denote the vector $(\alpha_1, \ldots, \alpha_J)$. The values α_j^* minimizing the sum of the squared weights of evidence can thus be found by solving the following minimization problem

$$\min f(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \left(\widehat{\beta}_j \phi_j(x_i) + \alpha_j\right)^2 \tag{32}$$

subject to

$$\sum_{j=1}^{J} \alpha_j = \widehat{\beta}_0. \tag{33}$$

Developing the square in (32), we get

$$f(\boldsymbol{\alpha}) = \sum_{j=1}^{J} \widehat{\beta}_j^2 \left(\sum_{i=1}^{n} \phi_j(x_i)^2\right) + n \sum_{j=1}^{J} \alpha_j^2 + 2 \sum_{j=1}^{J} \widehat{\beta}_j \alpha_j \sum_{i=1}^{n} \phi_j(x_i). \tag{34}$$

The first term in the right-hand side of (34) does not depend on $\boldsymbol{\alpha}$, and the third term vanishes when the J features are centered, i.e., when $\sum_{i=1}^n \phi_j(x_i) = 0$ for all j. Let us first assume that this condition is met. Then, we just need to minimize $\sum_{j=1}^J \alpha_j^2$ subject to (33). The solution is

$$\alpha_j^* = \widehat{\beta}_0 / J, \quad j = 1, \ldots, J. \tag{35}$$

In the case of logistic regression, where $\phi_j(x) = x_j$, the condition $\sum_{i=1}^n \phi_j(x_i) = 0$ can easily be ensured by centering the data before estimating the parameters. In the nonlinear case, the features ϕ_j are constructed during the learning process and they cannot be centered beforehand. Let μ_j denote the mean of feature ϕ_j, $\mu_j = \frac{1}{n}\phi_j(x_i)$, and $\phi'(x_i) = \phi(x_i) - \mu_j$ the centered feature values. We can write

$$w_{ij} = \beta_j \phi_j(x_i) + \alpha_j = \beta_j \phi_j'(x_i) + \alpha_j',$$

with $\alpha_j' = \alpha_j + \beta_j \mu_j$, and

$$\sum_{j=1}^{J} w_{ij} = \sum_{j=1}^{J} \beta_j \phi_j(x_i) + \beta_0 = \sum_{j=1}^{J} \beta_j \phi_j'(x_i) + \beta_0'$$

with $\beta_0' = \beta_0 + \sum_{j=1}^J \beta_j \mu_j$. As shown above, the optimal value of α_j' is

$$\alpha_j'^* = \frac{\widehat{\beta}_0'}{J} = \frac{\widehat{\beta}_0}{J} + \frac{1}{J} \sum_{j=1}^{J} \widehat{\beta}_j \mu_j.$$

Consequently, the optimal value of α_j is

$$\alpha_j^* = \alpha_j'^* - \widehat{\beta}_j \mu_j = \frac{\widehat{\beta}_0}{J} + \frac{1}{J} \sum_{q=1}^{J} \widehat{\beta}_q \mu_q - \widehat{\beta}_j \mu_j.$$

Remark 1. In this section, we have started from parameter estimates $\widehat{\beta}_j$, $j = 0, \ldots, J$ to compute the values α_j^* that give us the least informative mass functions, in terms of the sum of squared weights of evidence. We thus have a two-step process, where coefficients β_j are first estimated, and the α_j are determined in a second step. As a complementary approach, we can attempt to minimize the squared weights of evidence in the course of the learning process. In the simple case where the features are centered, the sum of squared weights of evidence has the following form, from (34) and (35):

$$\sum_{j=1}^{J} \beta_j^2 \left(\sum_{i=1}^{n} \phi_j(x_i)^2\right) + \frac{n}{J} \beta_0^2.$$

As a heuristic, we can add to the loss function a term $\lambda_1 \sum_{j=1}^J \beta_j^2 + \lambda_2 \beta_0^2$. We recognize the idea of *ridge regression* and ℓ_2-regularization, or *weight decay*. We can thus reinterpret regularization in the last layer of a neural network as a heuristic for minimizing the sum of squared weights of evidence, in application of the Least Commitment Principle. This remark also applies to the multi-category case addressed in the next section.

4.2. Multi-category case

In the multi-category case, we must determine both sets of coefficients $\{\beta_{jk}\}$ and $\{\alpha_{jk}\}$. As before, let $\widehat{\beta}_{jk}$ denote the maximum likelihood estimates of the weights β_{jk}, and let $\boldsymbol{\alpha}$ denote the vector of parameters α_{jk}. Any set of coefficients $\beta_{jk}^* = \widehat{\beta}_{jk} + c_j$ will produce the same output probabilities (30) as $\widehat{\beta}_{jk}$. The optimal parameter values β_{jk}^* and α_{jk}^* can, thus, be found by solving the following minimization problem

$$\min f(\boldsymbol{c}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K} \left[(\widehat{\beta}_{jk} + c_j)\phi_j(x_i) + \alpha_{jk}\right]^2 \tag{36}$$

subject to the K linear constraints

$$\sum_{j=1}^{J} \alpha_{jk} = \widehat{\beta}_{0k} + c_0, \quad k = 1, \ldots, K. \tag{37}$$

Proposition 2. *The solution of the minimization problem (36)–(37) is given by*

$$\beta_{0k}^* = \widehat{\beta}_{0k} - \frac{1}{K} \sum_{l=1}^{K} \widehat{\beta}_{0l}$$

and

$$\alpha_{jk}^* = \frac{1}{J} \left(\beta_{0k}^* + \sum_{j=1}^{J} \beta_{jk}^* \mu_j\right) - \beta_{jk}^* \mu_j. \tag{38}$$

Proof. See Appendix B.

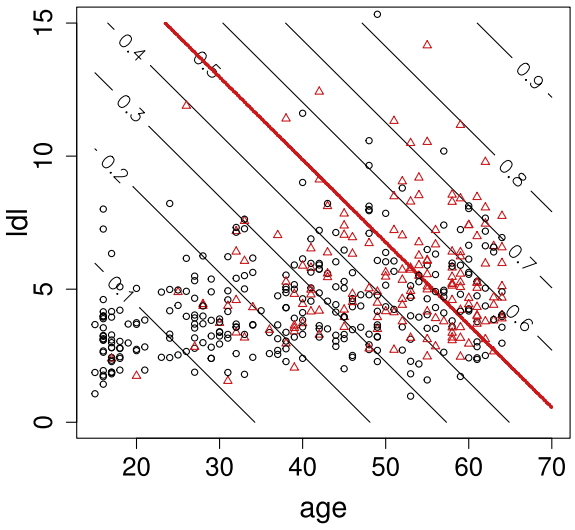


Fig. 5. Heart disease data, with the decision boundary (thick solid line) and the lines of equal positive class posterior probability for the logistic regression classifier. The positive and negative instances are identified by triangles and circles, respectively.

To get the least committed mass function m_i^* with minimum sum of squared weights of evidence and verifying (30), we thus need to center the rows of the $(J+1)\times K$ matrix $B=(\widehat{\beta}_{jk})$, set α_{jk}^* according to (38), and compute the weights of evidence w_k^- and w_k^+ from (25) and (27).

5. Numerical experiments

In this section, we illustrate through examples some properties of the mass functions computed by GLR classifiers. We demonstrate their use to interpret the computations performed by such networks, and to quantify prediction uncertainty. We start with a binary classification problem and logistic regression in Section 5.1. We then proceed with a multi-category dataset and a neural network model in Section 5.2.

5.1. Heart disease data

As an example of a real dataset, we considered the Heart Disease data[3] used in [44]. These data were collected as part of a study aiming to establish the intensity of ischemic heart disease risk factors in a high-incidence region in South Africa. The data represent white males between 15 and 64, and the response variable is the presence or absence of myocardial infarction (MI) at the time of the survey. There are 160 positive cases in this data, and a sample of 302 negative cases (controls). For display purposes, we considered only two input variables: age and low-density lipoprotein (LDL) cholesterol. The output variable Y takes values θ_1 and θ_2 for presence and absence of MI, respectively.

5.1.1. Analysis and interpretation of mass functions

Fig. 5 shows the data, with the decision boundary and the lines of equal class θ_1 posterior probability for the logistic regression classifier. The weights of evidence w_j as functions of x_j for the two input variables are shown in Fig. 6. We can see that an age greater than $\xi_1\approx 50$ is evidence for the presence of MI ($w_j>0$), whereas an age less than 50 is evidence for the absence of MI ($w_j<0$). For LDL, the cut-off point is $\xi_2\approx 6.87$. The corresponding mass functions m_j for each of the two features are displayed in Fig. 7. At the cut-off point ξ_j, the mass function m_j is vacuous, which indicates that feature x_j does not support any of the two classes.

Different views of the output mass functions m obtained after combining the two feature-based mass functions m_j, $j=1,2$ are shown in Fig. 8. We can see that there is no support for the positive class when both variables are below their cut-off points (lower-left part of Fig. 8(a)), whereas the positive class is fully plausible (i.e., there is no support for the negative class) when both variables are above their cut-off points (upper-right Fig. 8(b)). When both variables are close to their cut-off points, the ignorance $m(\Theta)$ is high (Fig. 8(c)). The conflict between the two feature mass functions m_1 and m_2 is high when the two pieces of evidence point two different hypotheses as it is the case, for instance, for a young subject with a high LDL level (upper-left corner of Fig. 8(c)). We can see that the DS perspective allows us to distinguish between lack of support, and conflicting evidence. In the classic probabilistic setting, both cases result in posteriori probabilities close to 0.5, as shown in Fig. 5. Information about the nature of the evidence that gave rise to the posterior class probabilities is lost when normalizing the contour function.

5.1.2. Decision analysis

With 0–1 losses, the pessimistic (maximum belief) and optimistic (MP) decisions rules based on output mass functions yield the same results as the decision rule based on output probabilities because, from Eqs. (20) and (22),

$$p(\theta_1)\geq p(\theta_2)\Leftrightarrow Bel(\{\theta_1\})\geq Bel(\{\theta_2\})\Leftrightarrow Pl(\{\theta_1\})\geq Pl(\{\theta_2\}).$$

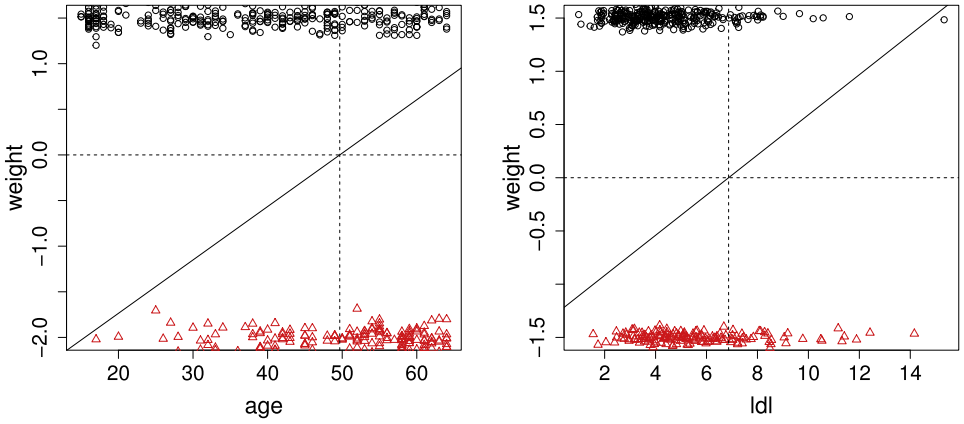


Fig. 6. Weights w_j as a function of x_j for variables age (left) and LDL (right). The feature values for positive and negative instances are shown, respectively, on the lower and upper horizontal axes, with some random vertical jitter to avoid overlap.

[3] This dataset can be downloaded from https://web.stanford.edu/~hastie/ElemStatLearn/.

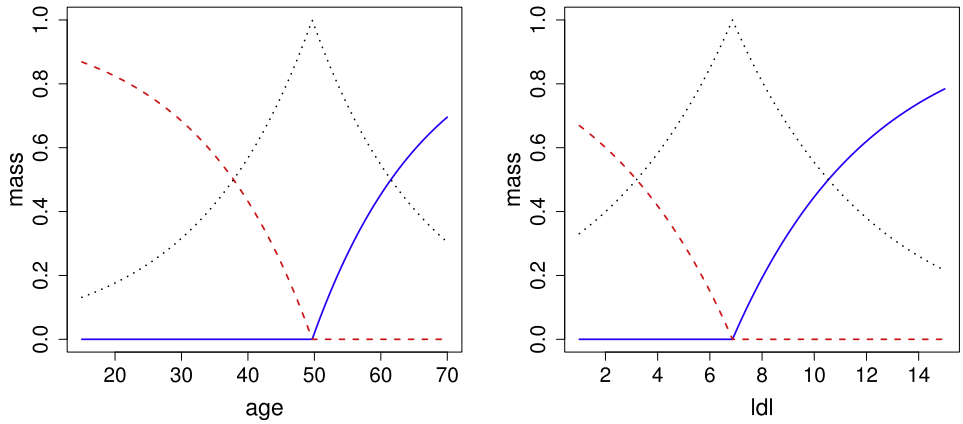


Fig. 7. Mass functions m_j for variables age (left) and LDL (right). The solid, broken and dotted lines correspond, respectively, to $m_j(\{\theta_1\})$, $m_j(\{\theta_2\})$ and $m_j(\Theta)$.

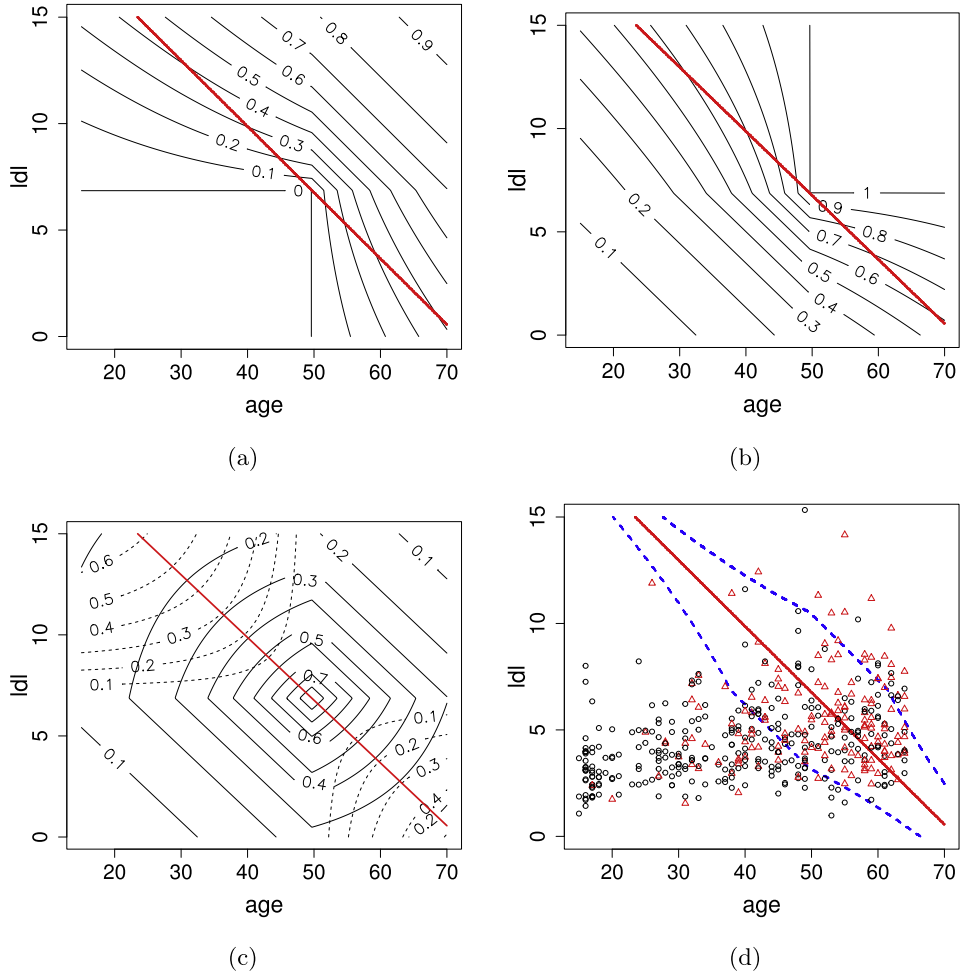


Fig. 8. (a): Curves of equal degree of belief $Bel(\{\theta_1\}) = m(\{\theta_1\})$ for the positive class; (b): curves of equal plausibility $pl(\theta_1)$; (c): ignorance $m(\Theta)$ (solid lines) and degree of conflict (broken lines); (d): Decision boundaries for the MP rule (solid line) and the ID rule (broken lines).

The corresponding decision boundary is shown as a solid line in Fig. 8(d). In contrast, the ID rule (recalled in Section 2.1.7) leads to the decision regions delimited by broken lines in Fig. 8(d). In the central region between the two curves, the intervals $[Bel(\{\theta_1\}), Pl(\{\theta_1\})]$ and $[Bel(\{\theta_2\}), Pl(\{\theta_2\})]$ are overlapping: consequently, the decision is $\{\theta_1, \theta_2\}$, i.e., there is not enough evidence to support selecting any of the two classes. Tables 1 and 2 show, respectively, the confusion matrices for the MP and ID rules, estimated by 10-fold cross-validation. (The results shown are averages over 30 replications of 10-fold cross-validation). The numbers in Tables 1 and 2 are expressed in percent and sum to 100. For instance, we can see from Table 1 that, on average, 13% of the data were in the positive class and were correctly classified by the MP rule, while 10.6% of the data were in the negative class and were wrongly classified by the same rule. The MP rule had an estimated error rate of $20.8 + 10.6 = 31.4\%$, while the ID rule had

Table 1
Confusion matrix for the MP rule, in % (Heart data).

		True class	
		Positive (θ_1)	Negative (θ_2)
Predicted	Positive (θ_1)	13.8	10.6
	Negative (θ_2)	20.8	54.8

Table 2
Confusion matrix for the ID rule, in % (Heart data).

		True class	
		Positive (θ_1)	Negative (θ_2)
Predicted	Positive (θ_1)	3.6	2.2
	Negative (θ_2)	9.8	42.2
	$\{\theta_1, \theta_2\}$	21.2	21.0

an error rate of $2.2 + 9.8 = 12.0\%$ and a rejection rate of 42.2%. If the rejected instances were classified randomly, the mean error rate would be $12 + 42.2/2 = 33.1\%$, which is only slightly higher than the MP error rate. This means that the ID rule is not overly cautious: it rejects instances that could hardly be classified better than randomly.

5.2. Gaussian multi-category data

As an example of a multi-category classification task with nonlinear decision boundaries, we consider an artificial dataset with $d = 2$ features, $K = 3$ equiprobable classes, and Gaussian class-conditional densities: $X|Y = \theta_k \sim \mathcal{N}(\mu_k, \Sigma_k)$, with

$$\mu_1 = \mu_2 = (0, 0)^T, \quad \mu_3 = (1, -1)^T$$

$$\Sigma_1 = 0.1I, \quad \Sigma_2 = 0.5I, \quad \Sigma_3 = \begin{pmatrix} 0.3 & -0.15 \\ -0.15 & 0.3 \end{pmatrix},$$

where I is the 2×2 identity matrix. We generated a learning set of size $n = 900$, and we trained a neural network with two layers of 20 and 10 rectified linear units (ReLU) [41]. The output layer had a softmax activation function. The network was trained in batch mode with a mini-batch size of 100. The first hidden layer had a drop-out rate [41] in the first hidden layer fixed to the standard value of 0.5. The weights between the last hidden layer and the output layer were penalized with an L_2 regularizer and a coefficient $\lambda = 0.5$ determined by 10-fold cross-validation.

5.2.1. Mass functions

Fig. 9 shows the data and the Bayes decision boundary. Contour lines of the masses assigned to different focal sets are shown in Fig. 10. We can see that masses are assigned to singletons in region of high class density, and to sets of classes in regions where these classes overlap. The output mass function m is the orthogonal sum of mass functions m_j provided by the 10 units in the last hidden layer. Plotting these mass functions allows us to interpret the role played by each of the hidden units. For instance, Fig. 11 shows the masses $m_j(A)$ assigned to different focal sets $A \subseteq \Theta$ by one of the hidden units. When the hidden unit output ϕ_j is small, the mass is distributed between $\{\theta_1\}$, $\{\theta_1, \theta_2\}$ and $\{\theta_1, \theta_3\}$. When ϕ_j is large, it supports $\{\theta_2\}$, $\{\theta_3\}$ and $\{\theta_2, \theta_3\}$.

5.2.2. Decision boundaries

The decision boundaries for the MP and ID rules are displayed in Fig. 12. We can see that the ID rule divides the feature space into six decision regions, corresponding to precise assignment to each of the three classes, and to imprecise assignment to subsets $\{\theta_1, \theta_2\}$, $\{\theta_2, \theta_3\}$ and $\Theta = \{\theta_1, \theta_2, \theta_3\}$. The existence of these "ambiguous" decisions is due to lack of evidence in regions where the classes overlap. We observe that regions corresponding to sets of classes partially include the Bayes boundary: the Bayes optimal decision, thus, often belongs to the set of decisions prescribed by the ID rule, including cases where the MP rule differs from the Bayes decision.

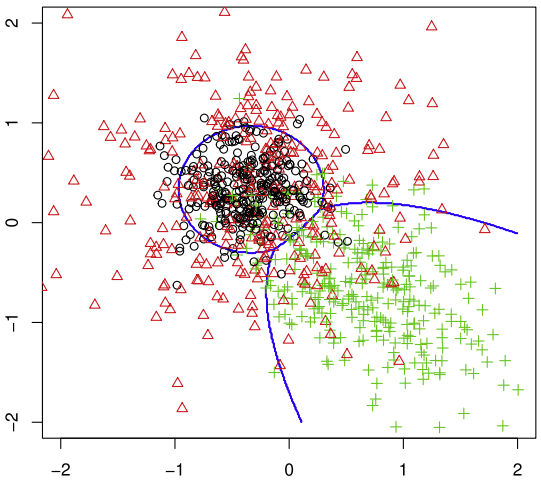


Fig. 9. Simulated Gaussian data with the Bayes decision boundary.

Table 3
Confusion matrix for the MP rule, in % (Gaussian data).

		True class		
		θ_1	θ_2	θ_3
Predicted	θ_1	26.8	9.7	2.6
	θ_2	5.6	18.2	1.5
	θ_3	0.9	5.4	29.2

Table 4
Confusion matrix for the ID rule, in % (Gaussian data).

		True class		
		θ_1	θ_2	θ_3
Predicted	θ_1	21.5	6.6	1.8
	θ_2	2.5	14.1	0.5
	θ_3	0.9	5.2	28.3
	$\{\theta_1, \theta_2\}$	6.8	5.3	1.0
	$\{\theta_2, \theta_3\}$	0.3	1.4	1.2
	$\{\theta_1, \theta_2, \theta_3\}$	0.9	0.8	1.0

5.2.3. Error rates

To estimate error rates, we generated a test dataset of size $n_t = 15{,}000$. The estimated Bayes error rate was 24.6%, and the estimated error rate of the MP rule was 25.7%. The confusion matrices for the MP and ID rules are shown, respectively, in Tables 3 and 4. The error rate of the ID rule is 17.5%, less that the Bayes error rate. Of course, this is compensated by assigning 16.46% of instances to a pair of classes, and 2.83% to the set of three classes. If one chooses a single class in each decision set randomly, the mean error rate will be $17.5 + 16.46/2 + 2.83/3 \approx 26.7\%$, which is only slightly higher than the MP error rate. This result suggests that the neural network classifier indeed does not perform much better than chance when the ID rule does not select a single class.

6. Conclusion

In this paper, we have revisited logistic regression and its extensions, including multilayer feedforward neural networks, by

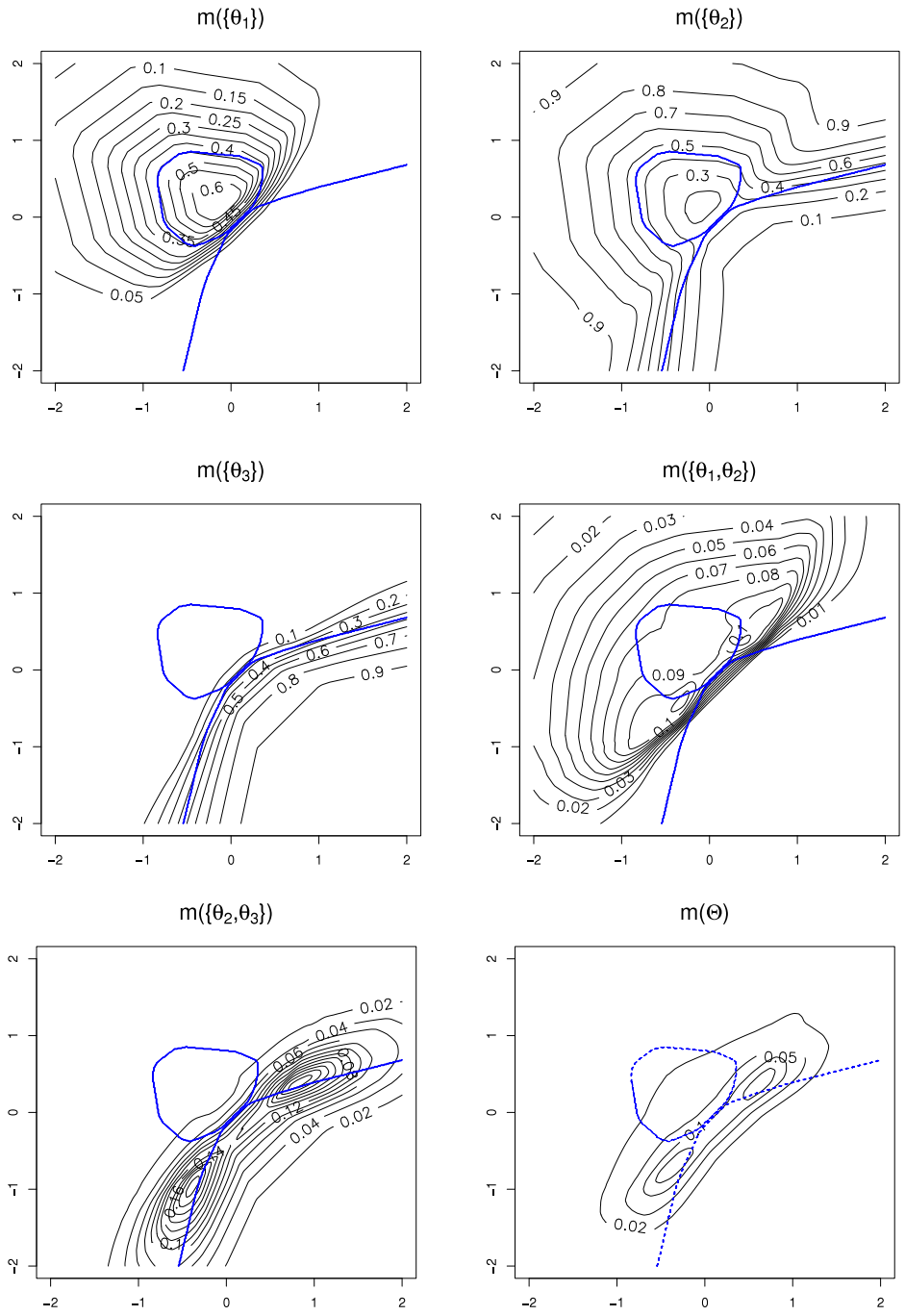


Fig. 10. Level curves of the output masses assigned to different focal sets (Gaussian data), with the MP decision boundary.

showing that these classifiers can be seen as converting (input or higher-level) features into mass functions and aggregating them by Dempster's rule of combination. The probabilistic outputs of these classifiers are the normalized plausibilities corresponding to the underlying combined mass function. This mass function has more degrees of freedom that the output probability distribution, and we have shown that it carries useful information. In particular, it makes it possible to distinguish between lack of evidence (when none of the features provides discriminant information) from conflicting evidence (when different features support different classes). This expressivity of mass functions allows us to gain insight into the role played by each input feature in logistic regression, and to interpret hidden unit outputs in multilayer neural networks. It also makes it possible to use decision rules, such as the interval dominance rule, which select a set of classes when the available evidence does not unambiguously point to a single class, thus reducing the error rate.

The significance of this result stems, in our view, from the fact that it sets DS theory as a suitable framework for analyzing and designing a wide range classifiers, including the now popular deep neural networks. Even though a lot of work has been done over the years applying belief functions to classification, this approach remained marginal in the vast landscape of statistical pattern recognition and machine learning techniques. The results presented in this paper show that belief functions are, in fact, ubiquitous in a large number of machine learning algorithms,

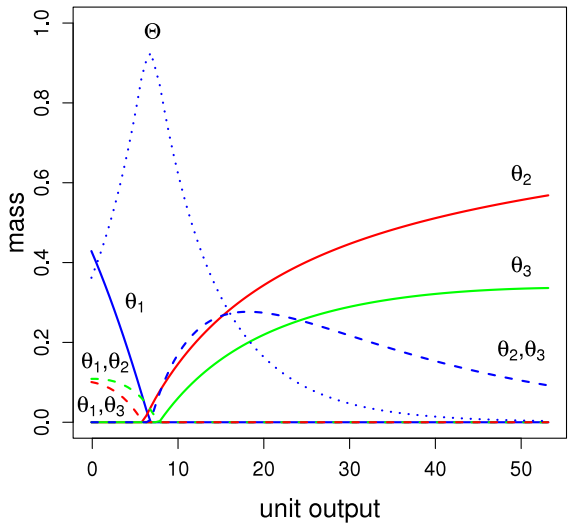


Fig. 11. Masses computed by a hidden unit, as functions of the unit outputs.

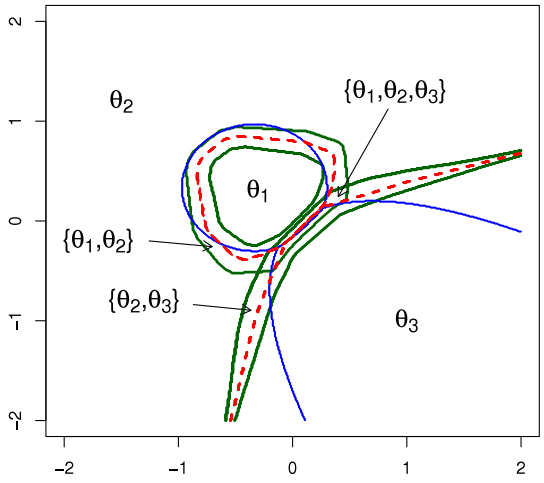


Fig. 12. Bayes decision boundary (thin solid line) and boundaries of the MP rule (thick broken line) and ID rule (thick solid line). The ID rule divides the feature space into six regions, with corresponding interpretations shown in the figure.

although this fact has been completely overlooked so far. This change of perspective opens the way to a whole research program, whose general objective is to better use existing classifiers and to design new models, based on the strong connection between GLR classifiers and DS theory laid bare in this paper. For instance, it would be interesting to study the properties of other decision rules in the belief function and imprecise probability frameworks, such as maximality and e-admissibility [37]. New classifier fusion schemes could be devised by combining the classifier output mass functions instead of aggregating decisions by majority voting or averaging probabilities. And alternatives to Dempster's rule, such as the cautious rule [29], could be investigated, to combine both feature-level mass functions inside the classifier, and output mass functions from a classifier ensemble.

Acknowledgments

This research was supported by the Labex MS2T, which was funded by the French Government, through the program "Investments for the future" by the National Agency for Research, France (reference ANR-11-IDEX-0004-02).

Appendix A. Proof of Proposition 1

A.1. Expression of m^+

As all positive masses m_k^+ defined by (26a) have the singletons $\{\theta_k\}$ and Θ as only focal elements, so has their orthogonal sum m^+. We thus have

$$m^+(\{\theta_k\}) \propto \left[1-\exp\left(-w_k^+\right)\right]\prod_{l\neq k}\exp\left(-w_l^+\right) =$$

$$\prod_{l\neq k}\exp\left(-w_l^+\right) - \prod_{l=1}^{K}\exp\left(-w_l^+\right) = \left[\exp(w_k^+)-1\right]\exp\left(-\sum_{l=1}^{K} w_l^+\right)$$

and $m^+(\Theta) \propto \exp\left(-\sum_{l=1}^{K} w_l^+\right)$. Consequently,

$$\sum_{k=1}^{K} m^+(\{\theta_k\}) + m^+(\Theta) \propto \exp\left(-\sum_{l=1}^{K} w_l^+\right)\left[\sum_{k=1}^{K}\exp(w_k^+) - K + 1\right]$$

and we have

$$m^+(\{\theta_k\}) = \frac{\exp(w_k^+)-1}{\sum_{l=1}^{K}\exp(w_l^+) - K + 1},\quad k=1,\ldots,K \tag{A.1a}$$

$$m^+(\Theta) = \frac{1}{\sum_{l=1}^{K}\exp(w_l^+) - K + 1}. \tag{A.1b}$$

We note that $m^+(\{\theta_k\})$ is an increasing function of the total weight of evidence w_k^+ supporting θ_k, and $m^+(\Theta)$ tends to one when all the positive weights w_k^+ tend to zero.

A.2. Expression of m^-

The degree of conflict when combining the negative mass functions m_k^-, $k=1,\ldots,K$, defined by (26b) is

$$\kappa^- = \prod_{k=1}^{K}\left[1-\exp(-w_k^-)\right]. \tag{A.2}$$

We thus have, for any strict subset $A \subset \Theta$,

$$m^-(A) = \frac{\left\{\prod_{\theta_k\notin A}\left[1-\exp(-w_k^-)\right]\right\}\left\{\prod_{\theta_k\in A}\exp(-w_k^-)\right\}}{1-\prod_{k=1}^{K}\left[1-\exp(-w_k^-)\right]}, \tag{A.3a}$$

and

$$m^-(\Theta) = \frac{\exp\left(-\sum_{k=1}^{K} w_k^-\right)}{1-\prod_{k=1}^{K}\left[1-\exp(-w_k^-)\right]}. \tag{A.3b}$$

From (28) and (A.2), the corresponding contour function is

$$pl^-(\theta_k) = \frac{\exp(-w_k^-)}{1-\prod_{l=1}^{K}\left[1-\exp(-w_l^-)\right]},\quad k=1,\ldots,K. \tag{A.4}$$

A.3. Combination of m^+ and m^-

Let $\eta^+ = \left(\sum_{l=1}^{K}\exp(w_l^+) - K + 1\right)^{-1}$ and $\eta^- = \left(1-\prod_{l=1}^{K}\left[1-\exp(-w_l^-)\right]\right)^{-1}$. From (A.1) and (A.3), the degree of conflict

between m^- and m^+ is

$$\kappa = \sum_{k=1}^{K} \left\{ m^+(\{\theta_k\}) \sum_{A \not\ni \theta_k} m^-(A) \right\}$$
$$= \sum_{k=1}^{K} \left\{ m^+(\{\theta_k\})(1 - pl^-(\theta_k)) \right\}$$
$$= \sum_{k=1}^{K} \left\{ \eta^+ \left(\exp(w_k^+) - 1 \right) \left[1 - \eta^- \exp(-w_k^-) \right] \right\}.$$

Let $\eta = (1-\kappa)^{-1}$. We have, for any $k \in \{1, \ldots, K\}$,

$$m(\{\theta_k\}) = \eta \left\{ m^+(\{\theta_k\}) \left[\sum_{A \ni \theta_k} m^-(A) \right] + m^-(\{\theta_k\}) m^+(\Theta) \right\} =$$
$$\eta \left\{ m^+(\{\theta_k\}) pl^-(\theta_k) + m^-(\{\theta_k\}) m^+(\Theta) \right\}.$$

Using Eqs. (A.1), (A.3), and (A.4), we get

$$m(\{\theta_k\}) = \eta \eta^- \eta^+ \exp(-w_k^-) \left\{ \exp(w_k^+) - 1 + \prod_{l \neq k} \left[1 - \exp(-w_l^-) \right] \right\}.$$

And for any $A \subseteq \Theta$ such that $|A| > 1$,

$$m(A) = \eta \; m^-(A) m^+(\Theta)$$
$$= \eta \; \eta^- \eta^+ \left\{ \prod_{\theta_k \notin A} \left[1 - \exp(-w_k^-) \right] \right\} \left\{ \prod_{\theta_k \in A} \exp(-w_k^-) \right\},$$

which completes the proof of Proposition 1.

Appendix B. Proof of Proposition 2

Developing the square in (36), we get

$$f(\mathbf{c}, \boldsymbol{\alpha}) = \sum_{j,k} (\widehat{\beta}_{jk} + c_j)^2 \left(\sum_{i=1}^{n} \phi_j(x_i)^2 \right) + n \sum_{j,k} \alpha_{jk}^2 + 2 \sum_{j,k} (\widehat{\beta}_{jk} + c_j) \alpha_{jk} \sum_{i=1}^{n} \phi_j(x_i). \tag{B.1}$$

Assuming, as in Section 4.1, the features ϕ_j to be centered, the last term in the right-hand side of (B.1) vanishes, and we get

$$f(\mathbf{c}, \boldsymbol{\alpha}) = \sum_{j=1}^{J} \left(\sum_{i=1}^{n} \phi_j(x_i)^2 \right) \sum_{k=1}^{K} (\widehat{\beta}_{jk} + c_j)^2 + n \sum_{j,k} \alpha_{jk}^2. \tag{B.2}$$

Due to constraints (37), for any c_0, the second term in the right-hand side of (B.2) is minimized for

$$\alpha_{jk} = \frac{1}{J} (\widehat{\beta}_{0k} + c_0), \quad \text{for } j = 1, \ldots, J \text{ and } k = 1 \ldots, K.$$

Hence, the problem becomes

$$\min_{\mathbf{c}} f(\mathbf{c}) = \sum_{j=1}^{J} \left(\sum_{i=1}^{n} \phi_j(x_i)^2 \right) \left\{ \sum_{k=1}^{K} (\widehat{\beta}_{jk} + c_j)^2 \right\} + \frac{n}{J} \sum_{k=1}^{K} (\widehat{\beta}_{0k} + c_0)^2.$$

Each of the $J+1$ terms in this sum can be minimized separately. The solution can easily be found to be

$$c_j^* = -\frac{1}{K} \sum_{k=1}^{K} \widehat{\beta}_{jk}, \quad j = 0, \ldots, J.$$

The optimum coefficients are, thus,

$$\beta_{jk}^* = \widehat{\beta}_{jk} - \frac{1}{K} \sum_{l=1}^{K} \widehat{\beta}_{jl}, \tag{B.3}$$

for $j = 0, \ldots, J$ and $k = 1 \ldots, K$, and $\alpha_{jk}^* = \beta_{0k}^*/J$ for $j = 1, \ldots, J$ and $k = 1 \ldots, K$.

Let us now consider the case where the features are not centered. As before, let $\phi_j'(x_i) = \phi_j(x_i) - \mu_j$ denote the centered feature values. We can write

$$w_{ijk} = \beta_{jk} \phi_j(x_i) + \alpha_{jk} = \beta_{jk} \phi_j'(x_i) + \alpha_{jk}',$$

with $\alpha_{jk}' = \alpha_{jk} + \beta_{jk} \mu_j$, and

$$\sum_{j=1}^{J} w_{ijk} = \sum_{j=1}^{J} \beta_{jk} \phi_j(x_i) + \beta_{0k} = \sum_{j=1}^{J} \beta_{jk} \phi_j'(x_i) + \beta_{0k}',$$

with $\beta_{0k}' = \beta_{0k} + \sum_{j=1}^{J} \beta_{jk} \mu_j$. The coefficients β_{jk} are not modified, except for $j = 0$. The optimal value of β_{0k} is

$$\beta_{0k}'^* = \widehat{\beta}_{0k}' - \frac{1}{K} \sum_{l=1}^{K} \widehat{\beta}_{0l}'$$
$$= \widehat{\beta}_{0k} + \sum_{j=1}^{J} \widehat{\beta}_{jk} \mu_j - \frac{1}{K} \sum_{l=1}^{K} \left(\widehat{\beta}_{0l} + \sum_{j=1}^{J} \widehat{\beta}_{jl} \mu_j \right)$$
$$= \widehat{\beta}_{0k} - \frac{1}{K} \sum_{l=1}^{K} \widehat{\beta}_{0l} + \sum_{j=1}^{J} \left(\widehat{\beta}_{jk} - \frac{1}{K} \sum_{l=1}^{K} \widehat{\beta}_{jl} \right) \mu_j$$
$$= \widehat{\beta}_{0k} - \frac{1}{K} \sum_{l=1}^{K} \widehat{\beta}_{0l} + \sum_{j=1}^{J} \beta_{jk}^* \mu_j.$$

Consequently,

$$\beta_{0k}^* = \beta_{0k}'^* - \sum_{j=1}^{J} \beta_{jk}^* \mu_j = \widehat{\beta}_{0k} - \frac{1}{K} \sum_{l=1}^{K} \widehat{\beta}_{0l}.$$

Now,

$$\alpha_{jk}'^* = \beta_{0k}'^*/J = \frac{1}{J} \left(\beta_{0k}^* + \sum_{j=1}^{J} \beta_{jk}^* \mu_j \right).$$

Hence,

$$\alpha_{jk}^* = \alpha_{jk}'^* - \beta_{jk}^* \mu_j = \frac{1}{J} \left(\beta_{0k}^* + \sum_{j=1}^{J} \beta_{jk}^* \mu_j \right) - \beta_{jk}^* \mu_j,$$

which completes the proof.

References

[1] A.P. Dempster, Upper and lower probabilities induced by a multivalued mapping, Ann. Math. Stat. 38 (1967) 325–339.

[2] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, Princeton, N.J., 1976.

[3] R.R. Yager, L. Liu (Eds.), Classic Works of the Dempster-Shafer Theory of Belief Functions, Springer, Heidelberg, 2008.

[4] L. Xu, A. Krzyzak, C.Y. Suen, Methods of combining multiple classifiers and their applications to handwriting recognition, IEEE Trans. Syst. Man Cybern. 22 (3) (1992) 418–435.

[5] G. Rogova, Combining the results of several neural network classifiers, Neural Netw. 7 (5) (1994) 777–781.

[6] B. Quost, M.-H. Masson, T. Denœux, Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules, Internat. J. Approx. Reason. 52 (3) (2011) 353–374.

[7] Y. Bi, The impact of diversity on the accuracy of evidential classifier ensembles, Int. J. Approx. Reason. 53 (4) (2012) 584–607.

[8] Z.-G. Liu, Q. Pan, J. Dezert, J.W. Han, Y. He, Classifier fusion with contextual reliability evaluation, IEEE Trans. Cybern. 48 (5) (2018) 1605–1618.
[9] H. Jiang, R. Wang, J. Gao, Z. Gao, X. Gao, Evidence fusion-based framework for condition evaluation of complex electromechanical system in process industry, Knowl.-Based Syst. 124 (2017) 176–187.
[10] P. Xu, F. Davoine, H. Zha, T. Denœux, Evidential calibration of binary SVM classifiers, Internat. J. Approx. Reason. 72 (2016) 55–70.
[11] P. Minary, F. Pichon, D. Mercier, E. Lefèvre, B. Droit, Face pixel detection using evidential calibration and fusion, Internat. J. Approx. Reason. 91 (2017) 202–215.
[12] Z.-G. Liu, Z. Zhang, Y. Liu, J. Dezert, Q. Pan, A new pattern classification improvement method with local quality matrix based on K-NN, Knowl.-Based Syst. 164 (2019) 336–347.
[13] T. Denœux, Analysis of evidence-theoretic decision rules for pattern classification, Pattern Recognit. 30 (7) (1997) 1095–1107.
[14] T. Denœux, P. Smets, Classification using belief functions: the relationship between the case-based and model-based approaches, IEEE Trans. Syst. Man Cybern. B 36 (6) (2006) 1395–1406.
[15] P. Smets, Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem, Internat. J. Approx. Reason. 9 (1993) 1–35.
[16] A. Appriou, Probabilités et incertitude en fusion de données multi-senseurs, Rev. Sci. Tech. Déf. 11 (1991) 27–40.
[17] Z.-G. Su, P.-H. Wang, Improved adaptive evidential k-NN rule and its application for monitoring level of coal powder filling in ball mill, J. Process Control 19 (10) (2009) 1751–1762.
[18] N. Guettari, A.S. Capelle-Laizé, P. Carré, Blind image steganalysis based on evidential k-nearest neighbors, in: 2016 IEEE International Conference on Image Processing, ICIP, 2016, pp. 2742–2746.
[19] X.-L. Chen, P.-H. Wang, Y.-S. Hao, M. Zhao, Evidential KNN-based condition monitoring and early warning method with applications in power plant, Neurocomputing (2018).
[20] T. Denœux, A *k*-nearest neighbor classification rule based on dempster-shafer theory, IEEE Trans. Syst. Man Cybern. 25 (05) (1995) 804–813.
[21] L. Jiao, Q. Pan, X. Feng, F. Yang, An evidential k-nearest neighbor classification method with weighted attributes, in: Proceedings of the 16th International Conference on Information Fusion, 2013, pp. 145–150.
[22] Z.-G. Liu, Q. Pan, J. Dezert, A new belief-based K-nearest neighbor classification method, Pattern Recognit. 46 (3) (2013) 834–844.
[23] C. Lian, S. Ruan, T. Denœux, An evidential classifier based on feature selection and two-step classification strategy, Pattern Recognit. 48 (2015) 2318–2327.
[24] C. Lian, S. Ruan, T. Denœux, Dissimilarity metric learning in the belief function framework, IEEE Trans. Fuzzy Syst. 24 (6) (2016) 1555–1564.
[25] Z.-G. Su, T. Denoeux, Y.-S. Hao, M. Zhao, Evidential K-NN classification with enhanced performance via optimizing a class of parametric conjunctive t-rules, Knowl.-Based Syst. 142 (2018) 7–16.
[26] T. Denœux, A neural network classifier based on dempster-shafer theory, IEEE Trans. Syst. Man Cybern. A 30 (2) (2000) 131–150.
[27] T. Denœux, Logistic regression revisited: belief function analysis, in: F. Cuzzolin, T. Denœux, S. Destercke, A. Martin (Eds.), Belief Functions: Theory and Applications: Fourth International Conference (BELIEF 2018), in: number 11069 in Lecture Notes in Artificial Intelligence, Springer, Compiègne, France, sept. 2018, pp. 57–64.
[28] Y.-C. Ko, H. Fujita, Evidential weights of multiple preferences for competitiveness, Inform. Sci. 354 (2016) 211–221.
[29] T. Denœux, Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence, Artificial Intelligence 172 (2008) 234–264.
[30] P. Smets, The canonical decomposition of a weighted belief, in: Int. Joint Conf. on Artificial Intelligence, San Mateo, Ca, Morgan Kaufman, 1995, pp. 1896–1901.
[31] F. Voorbraak, A computationally efficient approximation of Dempster-Shafer theory, Int. J. Man-Mach. Stud. 30 (1989) 525–536.
[32] B.R. Cobb, P.P. Shenoy, On the plausibility transformation method for translating belief function models to probability models, Internat. J. Approx. Reason. 41 (3) (2006) 314–330.
[33] G.J. Klir, M.J. Wierman, Uncertainty-based information, in: Elements of Generalized Information Theory, Springer-Verlag, New-York, 1999.
[34] D. Dubois, H. Prade, A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets, Int. J. Gen. Syst. 12 (3) (1986) 193–226.
[35] R.R. Yager, The entailment principle for Dempster-Shafer granules, Int. J. Intell. Syst. 1 (1986) 247–262.
[36] G. Shafer, Constructive probability, Synthese 48 (1) (1981) 1–60.
[37] M.C. Troffaes, Decision making under uncertainty using imprecise probabilities, Internat. J. Approx. Reason. 45 (1) (2007) 17–29.
[38] H. Jiang, Y. Dong, Structural regularization in quadratic logistic regression model, Knowl.-Based Syst. 163 (2019) 842–857.
[39] T.J. Hastie, R.J. Tibshirani, Generalized Additive Models, Chapman and Hall/CRC, London, 1990.
[40] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, in: D.E. Rumelhart, J. McClelland (Eds.), Parallel Distributed Processing, Vol. 1, MIT Press, Cambridge, MA, 1986, pp. 318–362.
[41] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, http://www.deeplearningbook.org.
[42] J. Moody, C.J. Darken, Fast learning in networks of locally-tuned processing units, Neural Comput. 1 (2) (1989) 281–294.
[43] B. Schölkopf, A. Smola, Learning with Kernels, MIT Press, 2002.
[44] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer, 2009.