



HAL
open science

Parametric classification with soft labels using the Evidential EM algorithm

Benjamin Quost, Thierry Denoeux, Shoumei Li

► **To cite this version:**

Benjamin Quost, Thierry Denoeux, Shoumei Li. Parametric classification with soft labels using the Evidential EM algorithm: Generative vs. discriminative models. 2017. hal-01525605v1

HAL Id: hal-01525605

<https://hal.science/hal-01525605v1>

Preprint submitted on 22 May 2017 (v1), last revised 5 Jul 2018 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Parametric classification with soft labels using the Evidential EM algorithm

Generative vs. discriminative models

Benjamin Quost · Thierry Denœux ·
Shoumei Li

Received: date / Accepted: date

Abstract Partially supervised learning extends both supervised and unsupervised learning, by considering situations in which only partial information about the response variable is available. In this paper, we consider partially supervised classification and we assume the learning instances to be labeled by Dempster-Shafer mass functions, called soft labels. Linear discriminant analysis and logistic regression are considered as special cases of generative and discriminative parametric models. We show that the evidential EM algorithm can be particularized to fit the parameters in each of these models. We describe experimental results with simulated data sets as well as with two real applications: K-complex detection in sleep EEGs signals and facial expression recognition. These results confirm the interest of using soft labels for classification as compared to potentially erroneous crisp labels, when the true class membership is partially unknown or ill-defined.

Keywords Partially supervised learning · Belief functions · Dempster-Shafer theory · Machine Learning · Uncertain data · Discriminant analysis · Logistic regression

Mathematics Subject Classification (2000) MSC 62H30 · MSC 62F86 · 68T10 · 68T37

This research was supported by a grant from the Beijing Government as part of the Overseas Talents program.

Benjamin Quost
Sorbonne Universités, Université de Technologie de Compiègne, CNRS, Heudiasyc (UMR 7253), France. E-mail: benjamin.quost@utc.fr

Thierry Denœux
Beijing University of Technology, College of Applied Sciences, Beijing, China and Sorbonne Universités, Université de Technologie de Compiègne, CNRS, Heudiasyc (UMR 7253), France. E-mail: thierry.denoeux@utc.fr

Shoumei Li
Beijing University of Technology, College of Applied Sciences, Beijing, China. E-mail: lisma@bjut.edu.cn

1 Introduction

Classically, a distinction is made in data analysis and machine learning between *supervised* and *unsupervised* learning. In supervised learning, the data set is composed of observations of the response variable and a list of input variables for n individuals of some population. The problem is then to predict the value of the response variable for a new individual. In unsupervised learning, no response variable is observed, and the task is to find some underlying structure in the data (such as a partition or a manifold).

In recent years, though, the distinction between supervised and unsupervised learning has been blurred by the introduction of new paradigms that lie between these two extremes. One of these paradigms is *semi-supervised* learning [2], in which the response variable is perfectly known for some individuals, and totally unknown for others. An even more general paradigm is *partially supervised learning* [9, 13, 19, 26, 4, 5], in which the values of the response variable for learning instances is only assumed to be partially known, i.e., is subject to some hard or soft constraints. Soft labels typically occur when no ground truth about class labels is available, and data have to be labeled by experts or using some indirect method (see, e.g. [3] for a concrete diagnosis application). In [9], the second author first considered a type of classification problem in which partial knowledge about class labels is expressed in the Dempster-Shafer framework by belief functions [7, 33]: we can then speak of *soft labels* [4]. The concept of soft label is particularly flexible, as it allows us to express uncertainty about class labels in a variety of ways, including sets, possibility distributions or probability distributions.

To learn from data with soft labels, nonparametric techniques such as the evidential k -nearest neighbor rule [9] and decision trees [12, 15, 35] have first been proposed. A general mechanism for parametric inference in the presence of uncertain data (of which soft labels are a special case) was later introduced in [10]. This mechanism, called the *Evidential EM (E^2M) algorithm*, extends the Expectation-Maximization (EM) algorithm [8] by allowing for maximum-likelihood estimation from *partially missing* data. The E^2M algorithm has been applied to a variety of models, including partially supervised Independent Factor Analysis [3], Hidden Markov Models with partially hidden states [30], fuzzy data clustering using Gaussian mixture models [29] and decision trees [23].

Two types of parametric models are commonly used in classification: *generative models* describe the joint distribution of the input vector W and the class label Z , while *discriminative models* represent the conditional distribution of Z given W . Generative models, such as Gaussian mixture models are suitable both for unsupervised learning (clustering) and for supervised classification. In contrast, discriminative models (such as logistic regression) cannot be used in an unsupervised setting, because they need observations of the response variable Y . However, they rely on fewer adaptive parameters and less restrictive assumptions, which may result in better predictive performance in supervised learning tasks, especially when class-conditional distributions are misspecified

[27][1, p. 204]. The question then arises of the relative performances of generative and discriminative models in the presence of soft labels. This question is investigated in this paper, with emphasis on linear classification. Partially supervised Linear Discriminant Analysis (LDA) and Logistic Regression (LR) based on the E²M algorithm will be introduced and compared using simulated and real data with soft labels. The objective of this paper is to study the influence of label uncertainty on the performances of these two classification models and, if possible, to formulate prescriptions regarding the suitability of each of these two approaches in situations where class labels are uncertain.

The rest of this paper is organized as follows. Background information about the theory of belief functions and the E²M algorithm will first be recalled in Section 2. Partially supervised LDA and LR will then be introduced in Section 3. The issues of performance evaluation and model selection will be addressed in Section 4, and experimental results will be presented in Section 5. Section 6 will conclude the paper.

2 Background

In order to make the paper self-contained, the basic concepts of Dempster-Shafer theory will first be recalled in Section 2.1. The notion of evidential likelihood and the E²M algorithm will then be described in Sections 2.2 and Sections 2.3, respectively.

2.1 Dempster-Shafer theory

In Dempster-Shafer theory [33], uncertain knowledge about some variable X taking values in a finite set \mathcal{X} is described by a *mass function*, which is defined a mapping m from the power set $2^{\mathcal{X}}$ to $[0, 1]$ verifying

$$\sum_{A \subseteq \mathcal{X}} m(A) = 1$$

and $m(\emptyset) = 0$. The subsets A of \mathcal{X} such that $m(A) > 0$ are called the *focal sets* of m . A mass function m with a single focal set is said to be *logical*. Logical mass functions are in one-to-one correspondence with the subsets of \mathcal{X} . The mass function $m_{\mathcal{X}}$ such that $m_{\mathcal{X}}(\mathcal{X}) = 1$ is said to be *vacuous*; it represents total ignorance.

From a mass function m , we can compute a belief function and a plausibility function, defined, respectively, as

$$Bel(A) = \sum_{B \subseteq A} m(B)$$

and

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

for all $A \subseteq \mathcal{X}$. Functions Bel and Pl are, respectively, completely monotone and completely alternating capacities [33]. When all the focal sets of m are singletons, m is said to be *Bayesian*; Bel and Pl then boil down to the same probability measure. When m is *consonant*, i.e., when its focal sets are nested, then Pl is a possibility measure [36]: it verifies

$$Pl(A \cup B) = \max(Pl(A), Pl(B))$$

for all $A, B \subseteq \mathcal{X}$. The mapping $pl : \mathcal{X} \rightarrow [0, 1]$ defined by $pl(x) = Pl(\{x\})$ for any $x \in \mathcal{X}$ is called the *contour function* of m . When m is consonant, the following equalities hold,

$$Pl(A) = \max_{x \in A} pl(x)$$

for all $A \subseteq \mathcal{X}$, and $\max_{x \in \mathcal{X}} pl(x) = 1$. When m is Bayesian, pl is a probability mass function and $\sum_{x \in \mathcal{X}} pl(x) = 1$.

Given two mass function m_1 and m_2 , their *orthogonal sum* is defined as the mass function $m_1 \oplus m_2$ such that $(m_1 \oplus m_2)(\emptyset) = 0$ and

$$(m_1 \oplus m_2)(A) = \frac{1}{1 - \kappa} \sum_{B \cap C = A} m_1(B)m_2(C) \quad (1)$$

for all $A \subseteq \mathcal{X}$, $A \neq \emptyset$, where

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B)m_2(C) \quad (2)$$

is called the *degree of conflict* between m_1 and m_2 . The orthogonal sum of m_1 and m_2 is well-defined as long as they are not totally conflicting, i.e., their degree of conflict is strictly less than 1. Operation \oplus is called *Dempster's rule of combination*. The contour function $pl_1 \oplus pl_2$ of $m_1 \oplus m_2$ is proportional to the product of the contour functions pl_1 and pl_2 associated with m_1 and m_2 : we have

$$(pl_1 \oplus pl_2)(x) = \frac{pl_1(x)pl_2(x)}{1 - \kappa} \quad (3)$$

for all $x \in \mathcal{X}$. The orthogonal sum of a Bayesian mass function m_1 and an arbitrary mass function m_2 is Bayesian; the degree of conflict between m_1 and m_2 can then be obtained from the contour functions as

$$\kappa = 1 - \sum_{x \in \mathcal{X}} pl_1(x)pl_2(x). \quad (4)$$

2.2 Evidential likelihood

Let X be a discrete random vector with finite sample space \mathcal{X} and probability mass function $p_X(x; \theta)$ assumed to be known up to a parameter $\theta \in \Theta$. After a realization x of X has been observed, the (complete-data) likelihood function is the mapping from Θ to $[0, 1]$ defined by

$$L_c(\theta) = p_X(x; \theta), \quad \forall \theta \in \Theta. \quad (5)$$

Let us now assume that x is not observed precisely, but we collect some evidence about x . This evidence induces partial knowledge of x described by a mass function m on \mathcal{X} . The likelihood function (5) can then be generalized [10] to

$$L(\theta) = \sum_{A \subseteq \mathcal{X}} m(A) \sum_{x \in A} p_X(x; \theta), \quad \forall \theta \in \Theta, \quad (6)$$

Function $L(\theta)$ defined by (6) is called the *evidential likelihood function* induced by the uncertain data m . It must be emphasized here that the notion of evidential likelihood is a new concept. This concept is distinct from the classical one of likelihood, because mass function m is *not* a realization of a random element. See [10] for a detailed discussion on this issue, and [6] for alternative definitions of likelihood in the case of imprecise observations. When m is logical with $m(A) = 1$ for some $A \subseteq \mathcal{X}$, then $L(\theta)$ is simply the probability of the event $X \in A$. The evidential likelihood function then becomes identical to the likelihood function under the Coarsening at Random (CAR) assumption [17]. However, we do not assume in our model that A is generated at random.

By permuting the two summations in (6), we get another expression for $L(\theta)$, as

$$L(\theta) = \sum_{x \in \mathcal{X}} p_X(x; \theta) \sum_{A \ni x} m(A) = \sum_{x \in \mathcal{X}} p_X(x; \theta) pl(x), \quad (7)$$

where pl is the contour function associated to m . Comparing (7) and (4), we can see that $1 - L(\theta)$ equals the degree of conflict between the uncertain data m and the probability mass function $p(x; \theta)$. Maximizing $L(\theta; m)$ thus amounts to minimizing the conflict between the data and the model.

Equation (7) also reveals that $L(\theta; pl)$ can alternatively be viewed as the expectation of $pl(X)$,

$$L(\theta) = \mathbb{E}_\theta[pl(X)]. \quad (8)$$

In the special case where $X = (X_1, \dots, X_n)$ is an independent sample, and assuming that the contour function pl can be decomposed as

$$pl(x) = pl_1(x_1) \dots pl_n(x_n), \quad (9)$$

a property called *cognitive independence* by Shafer [33], (8) simplifies to

$$L(\theta) = \prod_{i=1}^n \mathbb{E}_\theta[pl_i(X_i)]. \quad (10)$$

Finally, we can remark that the normalized likelihood $pl(\theta) = L(\theta)/L(\hat{\theta})$, where $\hat{\theta}$ is a maximizer of $L(\theta)$, and it is assumed that $L(\hat{\theta}) < \infty$, can be interpreted as the contour function of a consonant belief function on Θ (or, equivalently, as a possibility distribution on Θ) [11]. Equation (8) and (10) can directly be extended to the case of continuous data.

2.3 E²M algorithm

The E²M algorithm introduced in [10] is a generalization of the EM algorithm [8], which allows one to maximize the evidential likelihood (6)-(8). Similarly to the EM algorithm, each iteration q of E²M is composed of two steps:

1. The E-step requires the combination of $p_X(x; \theta^{(q)})$, the probability mass function of X for the current estimate $\theta^{(q)}$ of θ , with the contour function pl . The result of this combination is a probability mass function $p_X(\cdot|pl; \theta^{(q)}) = p_X(\cdot; \theta^{(q)}) \oplus pl$, which can be computed as

$$p_X(x|pl; \theta^{(q)}) = \frac{p_X(x; \theta^{(q)})pl(x)}{L(\theta^{(q)}; pl)},$$

for all $x \in \mathcal{X}$. Then, the expectation of the complete-data log-likelihood $\ell(\theta) = \log p_X(X; \theta)$ with respect to $p(x|pl; \theta^{(q)})$ is calculated,

$$Q(\theta, \theta^{(q)}) = \frac{\sum_{x \in \mathcal{X}} \ell(\theta; x) p_X(x; \theta^{(q)}) pl(x)}{L(\theta^{(q)}; pl)}. \quad (11)$$

2. The M-step then consists in maximizing function $Q(\theta, \theta^{(q)})$ with respect to θ , finding the new estimate $\theta^{(q+1)}$ such that $Q(\theta^{(q+1)}, \theta^{(q)}) \geq Q(\theta, \theta^{(q)})$ for all $\theta \in \Theta$.

The E- and M-steps are iterated until $L(\theta^{(q+1)}) - L(\theta^{(q)}) \leq \epsilon$ for some arbitrarily small ϵ .

The E²M algorithm was shown in [10] to increase the evidential likelihood as each iteration, i.e., to ensure that $L(\theta^{(q+1)}) \geq L(\theta^{(q)})$ for all q . Consequently, it converges to a local maximum of the evidential likelihood if this function is bounded from above.

3 Application to linear classification models

In this section, the complete data are assumed to consist in an iid sample $X = \{(W_i, Z_i)\}_{i=1}^n$ from (W, Z) , where W is a d -dimensional random input vector and Z is the class variable, taking values in a finite set $\mathcal{Z} = \{1, \dots, K\}$. The complete dataset $x = \{(w_i, z_i)\}_{i=1}^n$ is a realization from X . Each input vector w_i is known, but z_i is only partially observed. We denote by $pl_{ik} = pl_i(k)$ the plausibility that instance i belongs to class k . We thus have a learning set $\{(w_i, pl_i)\}_{i=1}^n$ with soft labels. In the sequel, we will consider the application of the E²M algorithm to estimate the parameter of two models: LDA and LR using such data.

3.1 Linear Discriminant Analysis

LDA is based on the assumption that the conditional distribution of W given $Z = k$ is multivariate normal with mean μ_k and covariance matrix Σ independent on k :

$$W|(Z = k) \sim \mathcal{N}(\mu_k, \Sigma), \quad k = 1, \dots, K.$$

Let π_k be the marginal probability that $Z = k$, and

$$\theta = (\mu_1, \dots, \mu_K, \Sigma, \pi_1, \dots, \pi_{K-1})$$

the parameter vector. The complete-data likelihood is

$$L_c(\theta) = \prod_{i=1}^n p(w_i|Z_i = z_i)p(z_i) \quad (12a)$$

$$= \prod_{i=1}^n \prod_{k=1}^K \phi(w_i; \mu_k, \Sigma)^{z_{ik}} \pi_k^{z_{ik}}, \quad (12b)$$

where $\phi(\cdot; \mu_k, \Sigma)$ is the multivariate normal density,

$$\phi(w; \mu_k, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (w - \mu)^T \Sigma^{-1} (w - \mu) \right\},$$

and z_{ik} is a binary class indicator variable, such that $z_{ik} = 1$ if $z_i = k$ and $z_{ik} = 0$ otherwise.

Under the assumption of cognitive independence (9), the contour function on \mathcal{X} is $pl(x) = \prod_{i=1}^n pl_i(x_i)$, with

$$pl_i(x_i) = \begin{cases} pl_{ik} & \text{if } x_i = (w_i, k) \text{ for some } k = 1, \dots, K \\ 0 & \text{otherwise.} \end{cases}$$

The evidential likelihood (10) is thus

$$L(\theta) = \prod_{i=1}^n \sum_{k=1}^K pl_{ik} \phi(w_i; \mu_k, \Sigma) \pi_k. \quad (13)$$

We can remark that, when there is no uncertainty, i.e., when $pl_{ik} = z_{ik}$ for all (i, k) , we have

$$\sum_{k=1}^K pl_{ik} \phi(w_i; \mu_k, \Sigma) \pi_k = \prod_{k=1}^K \phi(w_i; \mu_k, \Sigma)^{pl_{ik}} \pi_k^{pl_{ik}},$$

and the evidential likelihood (13) becomes identical to the complete-data likelihood (12b). When, on the other hand, uncertainty is maximal, i.e., class labels are completely unknown, then $pl_{ik} = 1$ for all (i, k) , and the evidential likelihood (13) becomes

$$L(\theta) = \prod_{i=1}^n \sum_{k=1}^K \phi(w_i; \mu_k, \Sigma) \pi_k,$$

which is the likelihood function corresponding to the unsupervised case.

In the E-step of the E²M algorithm for this model, we compute the expectation of the complete-data log-likelihood

$$\ell_c(\theta) = \sum_{i=1}^n \sum_{k=1}^K z_{ik} [\log \phi(w_i; \mu_k, \Sigma) + \log \pi_k]$$

with respect to the combined probability mass function

$$p_X(x|pl; \theta^{(q)}) = \prod_{i=1}^n p(x_i|pl_i; \theta^{(q)}),$$

with

$$p(x_i|pl_i; \theta^{(q)}) = \begin{cases} \frac{pl_{ik}\pi_k^{(q)}\phi(w_i; \mu_k^{(q)}, \Sigma^{(q)})}{\sum_{\ell} pl_{i\ell}\pi_{\ell}^{(q)}\phi(w_i; \mu_{\ell}^{(q)}, \Sigma^{(q)})} & \text{if } x_i = (w_i, k) \text{ for some } k \\ 0 & \text{otherwise.} \end{cases}$$

We get

$$Q(\theta, \theta^{(q)}) = \sum_{i=1}^n \sum_{k=1}^K \zeta_{ik}^{(q)} [\log \phi(w_i; \mu_k, \Sigma)\pi_k + \log \pi_k], \quad (14)$$

with

$$\zeta_{ik}^{(q)} = \mathbb{E}(Z_{ik}|pl; \theta^{(q)}) = \frac{pl_{ik}\pi_k^{(q)}\phi(w_i; \mu_k^{(q)}, \Sigma^{(q)})}{\sum_{\ell} pl_{i\ell}\pi_{\ell}^{(q)}\phi(w_i; \mu_{\ell}^{(q)}, \Sigma^{(q)})}. \quad (15)$$

We can remark that the function $Q(\theta, \theta^{(q)})$ defined by (14) has exactly the same form as the function computed in the E-step of the EM algorithm applied to the normal mixture model in the unsupervised case (see, e.g., [25, pages 81-83]). When class labels are completely unknown, i.e., $pl_{ik} = 1$ for all i and k , then $\zeta_{ik}^{(q)}$ defined by (15) becomes equal to the conditional expectation of Z_{ik} given $W_i = w_i$, and E²M boils down to the classical EM algorithm.

Because of the formal similarity with the EM algorithm, the parameter values maximizing $Q(\theta, \theta^{(q)})$ can be readily obtained as

$$\pi_k^{(q+1)} = \frac{1}{n} \sum_{i=1}^n \zeta_{ik}^{(q)}, \quad \mu_k^{(q+1)} = \frac{\sum_{i=1}^n \zeta_{ik}^{(q)} w_i}{\sum_{i=1}^n \zeta_{ik}^{(q)}}, \quad (16a)$$

$$\Sigma^{(q+1)} = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \zeta_{ik}^{(q)} (w_i - \mu_k^{(q+1)})(w_i - \mu_k^{(q+1)})^T. \quad (16b)$$

3.2 Logistic regression

In contrast with LDA, LR starts with a model of the conditional distribution of Z given $W = w$. Specifically, assume that the conditional probabilities of each class given $W = w$ are given by

$$p_k(w; \theta) = \frac{\exp(\beta_k^T \tilde{w})}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_\ell^T \tilde{w})}, \quad k = 1, \dots, K-1 \quad (17a)$$

$$p_K(w; \theta) = \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_\ell^T \tilde{w})}, \quad (17b)$$

where $p_k(w; \theta)$ is a notation for $\Pr(Z = k | W = w; \theta)$, β_k is a $p+1$ -dimensional vector of coefficients, $\theta = (\beta_1^T, \dots, \beta_{K-1}^T)^T$ is the vector of all parameters in the model, and $\tilde{w} = (1, w^T)^T$ is an extended input vector. Logistic regression proceeds by maximizing the conditional likelihood

$$L_c(\theta) = \prod_{i=1}^n \Pr(Z_i = z_i | w_i; \theta) \quad (18a)$$

$$= \prod_{i=1}^n \prod_{k=1}^K p_k(w; \theta)^{z_{ik}}. \quad (18b)$$

After plugging the expression of the conditional class probabilities (17) into (18b), and taking the logarithm, we obtain the following expression for the complete-data conditional log-likelihood,

$$\ell_c(\theta) = \sum_{i=1}^n \left\{ \sum_{k=1}^{K-1} z_{ik} \beta_k^T \tilde{w}_i - \log \left(1 + \sum_{k=1}^{K-1} \exp(\beta_k^T \tilde{w}_i) \right) \right\}. \quad (19)$$

Under the same cognitive independence assumption (9) as before, the evidential likelihood is

$$L(\theta) = \prod_{i=1}^n \sum_{k=1}^K p_{l_{ik}} p_k(w_i; \theta). \quad (20)$$

We can easily check that $L(\theta) = L_c(\theta)$ whenever $p_{l_{ik}} = z_{ik}$ for all (i, k) , i.e., when there is no label uncertainty. On the other hand, in case of maximal uncertainty, i.e., when $p_{l_{ik}} = 1$ for all (i, k) , we have $L(\theta) = 1$ for all θ , and the model parameters can no longer be estimated.

The E²M algorithm for the maximization of the evidential likelihood (20) can be described as follows [28]. In the E-step, we compute the expectation of the complete-data log-likelihood (19) with respect to the combined probability mass function

$$p_Z(z | pl; \theta^{(q)}) = \prod_{i=1}^n p_{Z_i}(z_i | pl_i; \theta^{(q)}),$$

with

$$p_{Z_i}(k | pl_i; \theta^{(q)}) = \frac{p_{l_{ik}} p_k(w_i; \theta^{(q)})}{\sum_{\ell} p_{l_{i\ell}} p_\ell(w_i; \theta^{(q)})}, \quad k = 1, \dots, K.$$

We get

$$Q(\theta, \theta^{(q)}) = \sum_{i=1}^n \left\{ \sum_{k=1}^{K-1} \zeta_{ik}^{(q)} \beta_k^T \tilde{w}_i - \log \left(1 + \sum_{k=1}^{K-1} \beta_k^T \tilde{w}_i \right) \right\}, \quad (21)$$

with

$$\zeta_{ik}^{(q)} = \mathbb{E}(Z_{ik} | p_l; \theta^{(q)}) = \frac{p_{ik} p_k(w_i; \theta^{(q)})}{\sum_{\ell} p_{i\ell} p_{\ell}(w_i; \theta^{(q)})}. \quad (22)$$

The maximization of (21) cannot be performed in one step and requires an iterative optimization procedure, such as the Newton-Raphson algorithm. It is actually not necessary to maximize function $Q(\theta, \theta^{(q)})$: we may simply make a step uphill, i.e., find some new estimate $\theta^{(q+1)}$ such that $Q(\theta^{(q+1)}, \theta^{(q)}) > Q(\theta^{(q)}, \theta^{(q)})$. Such a procedure is classically called a *Generalized EM algorithm* [24, page 24]. An uphill step starting from the previous estimate $\theta^{(q)}$ can be made by carrying out one iteration of the Newton-Raphson algorithm with line search, i.e., by using the following update rule,

$$\theta^{(q+1)} = \theta^{(q)} - \eta \left[\frac{\partial^2 Q(\theta, \theta^{(q)})}{\partial \theta \partial \theta^T} \right]_{\theta=\theta^{(q)}}^{-1} \left. \frac{\partial Q(\theta, \theta^{(q)})}{\partial \theta} \right|_{\theta=\theta^{(q)}},$$

where η is the step size. Typically, η is initially set to 1 and, if the objective function has not increased, it is repeatedly cut by two until an uphill step is achieved. The gradient vector and Hessian matrix can be computed using matrix operations as in standard LR (see, e.g., [16, 22] for details).

4 Performance evaluation and model selection

Performance evaluation and model selection are fundamental issues when designing of classifiers. Typically, a loss function is defined, and the performance of a classifier is measured by its expected loss, or risk. If the loss is assumed to be equal to 1 for misclassification and 0 in case for correct classification, then the risk is the error probability. This probability may be estimated from an independent test set or using cross-validation. In the case of data with soft labels, error (or risk) estimation becomes more difficult, because the true class is uncertain. This issue is addressed in this section.

4.1 Lower and upper expected loss

Assume that, for some test pattern i with soft label m_i , we get the predicted class \hat{z}_i . Let $\lambda(\hat{z}_i, z_i)$ denote the loss incurred if the true class is z_i . The actual loss is unknown, because we only have partial knowledge of z_i , expressed by

mass function m_i . However, we can compute lower and upper expected values [7] as follows:

$$\lambda_*(\widehat{z}_i, m_i) = \sum_{A \subseteq \mathcal{Z}} m_i(A) \min_{k \in A} \lambda(\widehat{z}_i, k) \quad (23a)$$

$$\lambda^*(\widehat{z}_i, m_i) = \sum_{A \subseteq \mathcal{Z}} m_i(A) \max_{k \in A} \lambda(\widehat{z}_i, k). \quad (23b)$$

The lower and upper expected losses can be seen as, respectively, optimistic and pessimistic assessments of the unknown loss $\lambda(\widehat{z}_i, z_i)$. A trade-off between these two assessments can be achieved by computing a convex sum

$$\lambda_\rho(\widehat{z}_i, m_i) = \rho \lambda^*(\widehat{z}_i, m_i) + (1 - \rho) \lambda_*(\widehat{z}_i, m_i), \quad (24)$$

where $\rho \in [0, 1]$ is a pessimism index [20, 34].

In the special case where λ is the 0-1 loss function, defined by $\lambda(\widehat{z}_i, z_i) = I(\widehat{z}_i \neq z_i)$, where $I(\cdot)$ is the indicator notation, we have

$$\min_{k \in A} \lambda(\widehat{z}_i, k) = \begin{cases} 0 & \text{if } \widehat{z}_i \in A \\ 1 & \text{otherwise.} \end{cases}$$

Consequently,

$$\lambda_*(\widehat{z}_i, m_i) = \sum_{A \subseteq \mathcal{Z}: \widehat{z}_i \notin A} m_i(A) \quad (25a)$$

$$= 1 - \sum_{A \subseteq \mathcal{Z}: \widehat{z}_i \in A} m_i(A) \quad (25b)$$

$$= 1 - pl_i(\widehat{z}_i). \quad (25c)$$

Similarly,

$$\max_{k \in A} \lambda(\widehat{z}_i, k) = \begin{cases} 0 & \text{if } A = \{\widehat{z}_i\} \\ 1 & \text{otherwise.} \end{cases}$$

Hence,

$$\lambda^*(\widehat{z}_i, m_i) = 1 - m_i(\{\widehat{z}_i\}). \quad (26)$$

The width of the lower-upper expected loss interval reflects the imprecision of the mass function m_i . It is null if m_i is Bayesian, and it is equal to one if m_i is vacuous.

We can also remark that the lower expected 0-1 loss criterion (25) has an interesting alternative interpretation in terms of degree of conflict. Let \widehat{m}_i the *certain* mass function defined by $\widehat{m}_i(\{\widehat{z}_i\}) = 1$. The degree of conflict (2) between \widehat{m}_i and m_i is

$$\kappa_i = \sum_{B \cap C = \emptyset} \widehat{m}_i(B) m_i(C) \quad (27a)$$

$$= \sum_{C \not\ni \widehat{z}_i} m_i(C) \quad (27b)$$

$$= 1 - pl_i(\widehat{z}_i) = \lambda_*(\widehat{z}_i, m_i). \quad (27c)$$

The lower expected loss is thus equal to the degree of conflict between the prediction \hat{z}_i and the mass function m_i representing one's knowledge about the true class.

4.2 Application to performance assessment

Given a test set of size n_t , with soft labels $\{m_i\}_{i=1}^{n_t}$, lower and upper test error rates can be defined by averaging the lower and upper expected losses,

$$\underline{\lambda} = \frac{1}{n_t} \sum_{i=1}^{n_t} \lambda_*(\hat{z}_i, m_i) = 1 - \frac{1}{n_t} \sum_{i=1}^{n_t} pl_i(\hat{z}_i) \quad (28a)$$

$$\bar{\lambda} = \frac{1}{n_t} \sum_{i=1}^{n_t} \lambda^*(\hat{z}_i, m_i) = 1 - \frac{1}{n_t} \sum_{i=1}^{n_t} m_i(\{\hat{z}_i\}). \quad (28b)$$

They can be seen as, respectively, optimistic and pessimistic assessments of the test error rate. In the fully supervised case with certain mass function m_i verifying $m_i(\{z_i\}) = 1$, both $\underline{\lambda}$ and $\bar{\lambda}$ boil down to the usual test error rate. In the unsupervised case, we have $\underline{\lambda} = 0$ and $\bar{\lambda} = 1$, meaning that the error rate cannot be assessed.

Given a set of models (such as, e.g., LDA and LR with different subsets of input variables), their performances can be assessed using cross-validation estimates of the error rates. Typically, the learning set is partitioned into B blocks of approximately equal size, and B classifiers are trained, leaving one of the B blocks aside. The classifier performance is then assessed from its predictions on the block that was left aside. Cross-validation estimates $\underline{\lambda}_{CV}$ and $\bar{\lambda}^{CV}$ can be obtained in this way.

Given two classifiers C_1 and C_2 , and their lower and upper cross-validation error rates $\{\underline{\lambda}_{CV}^{(i)}, \bar{\lambda}_{CV}^{(i)}\}$, $i = 1, 2$, classifier C_1 can be considered to perform better than C_2 if $\bar{\lambda}_{CV}^{(1)} < \underline{\lambda}_{CV}^{(2)}$. This criterion yields only a partial order over classifiers. If a total order is needed, then we can base the comparison on the weighted means $(1 - \rho)\underline{\lambda}_{CV}^{(i)} + \rho\bar{\lambda}_{CV}^{(i)}$, after fixing the value of the pessimism index ρ .

5 Experiments

In this section, we report on experiments with LDA and LR trained from data with soft labels. In Section 5.1, we will first consider simulated and real data in which noise has been artificially introduced in class labels. We will then compare the performances of classifiers trained using the noisy labels, with those of classifiers trained using soft labels encoding labeling uncertainty. In Section 5.2, we will describe two real applications in which soft labels naturally arise.

Table 1: Real datasets used in Section 5.1.

dataset	amount of instances	number of classes	dimension
Breastcancer	569	2	30
Ionosphere	350	2	34
Iris	150	3	4
Satimage	6435	6	36
Sonar	208	2	60
Vowel	990	11	10
Wine	178	3	13

5.1 Simulations

Experimental procedure One synthetic dataset and seven real datasets are considered in this experiment. The synthetic dataset, *synthetic* with $n = 200$ instances was generated according to a Gaussian mixture model with $g = 3$ classes in proportions $\pi_1 = 0.45$, $\pi_2 = 0.35$ and $\pi_3 = 0.2$ in a $p = 2$ -dimensional space, and

$$\mu_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \mu_3 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The real datasets are classical datasets of the literature. Their main characteristics are reported in Table 1.

For each dataset, we considered two different classification models: LDA and LR. In order to evaluate the benefit of using soft class information in the learning process, we introduced noise in the data as follows. For each instance, a probability p_i was generated at random according to a beta distribution with fixed variance (set to 0.04) and expectation \bar{p} set to 0.1, 0.2, ..., 0.9 for the synthetic dataset, and 0.1, 0.2, ..., 0.7 for the others. Then, with probability p_i , the label was replaced with a label picked at random.

For each kind of classifier, we compared four kinds of labeling strategies: learning from the noisy labels, using “adaptive semi-supervised learning” (where a label z_i is kept if $p_i \leq 0.5$, the instance being considered as unlabeled otherwise), and using two kinds of soft labels: one where each p_i is taken into account, and one where the expected mislabeling rate is used instead. In the former case, if k^* is the observed label, $pl_{ik^*} = 1$ and $pl_{i\ell} = p_i$ for all $\ell \neq k^*$; the latter strategy replaces each p_i by the mean label noise \bar{p} . Less information is then used, but we might argue that this strategy is much easier to set up than the previous one, which requires an expert to assess each probability p_i of a label being erroneous.

For the synthetic dataset, we used a test set of size 1000 from the same distribution as the training data. For the real datasets, the classifiers were trained using 2/3 of the instances picked at random, the remaining ones being used for testing. When fitting LDA to the data, the “perfect” model (i.e., obtained

from the data without corrupted labels) was first determined, by considering ten different starting points (the initial proportions were set to $1/g$, the initial covariance matrices to the identity matrix, and the expectations vectors were sampled at random in the training set) and retaining the best solution (i.e., the one with the highest log-likelihood). The corresponding “actual” parameters were then used as starting values for the four LDA corresponding to the various labeling strategies mentioned above. For logistic regression, the initial parameter matrix was set to the $(p + 1) \times (g - 1)$ null matrix, and we used quadratic penalization with coefficient $\lambda = 0.01$. For each dataset, this whole procedure (training and test instance selection, introduction of label noise, model training, and model testing) was repeated 50 times, and average error rates and 95% confidence intervals were computed.

Results and discussion Overall, taking into account label uncertainty almost always significantly improves classification accuracy, as compared to noisy labels or the semi-supervised strategy. This is particularly true for LDA, where the benefit of using soft labels is critical when the level of label noise is high. For a majority of datasets, soft labels based on the expected mislabeling rate are almost as effective as those based on individual mislabeling rates (see, e.g., Figures 1(a), 1(b), 2(a)). For some datasets, however, there is a significant difference between both types of soft labels, the more informative one allowing for better performances (see Figures 3(b), 4(a)). The only case where using soft labels does not result in significant performance improvement with LDA is the Vowel dataset (Figure 4(a)). This exception can be explained by a large number of classes, a high overlap between classes, and perhaps a strong departure from normality, which makes learning with weak class information inefficient.

The benefit of using soft labels together with logistic regression also exists but it is less important than it is with LDA. Generally, soft labels make it possible to improve classification accuracy, but the difference with the adaptive semi-supervised strategy — and sometimes even noisy labels — is less remarkable, especially when mislabeling probability is high. The exceptions are the Satimage and Vowel datasets, where both strategies involving soft labels perform critically better than LDA (Figures 3(f) and 4(e)). Overall, LDA outperforms logistic regression for a majority of datasets, especially when soft labels are highly uncertain.

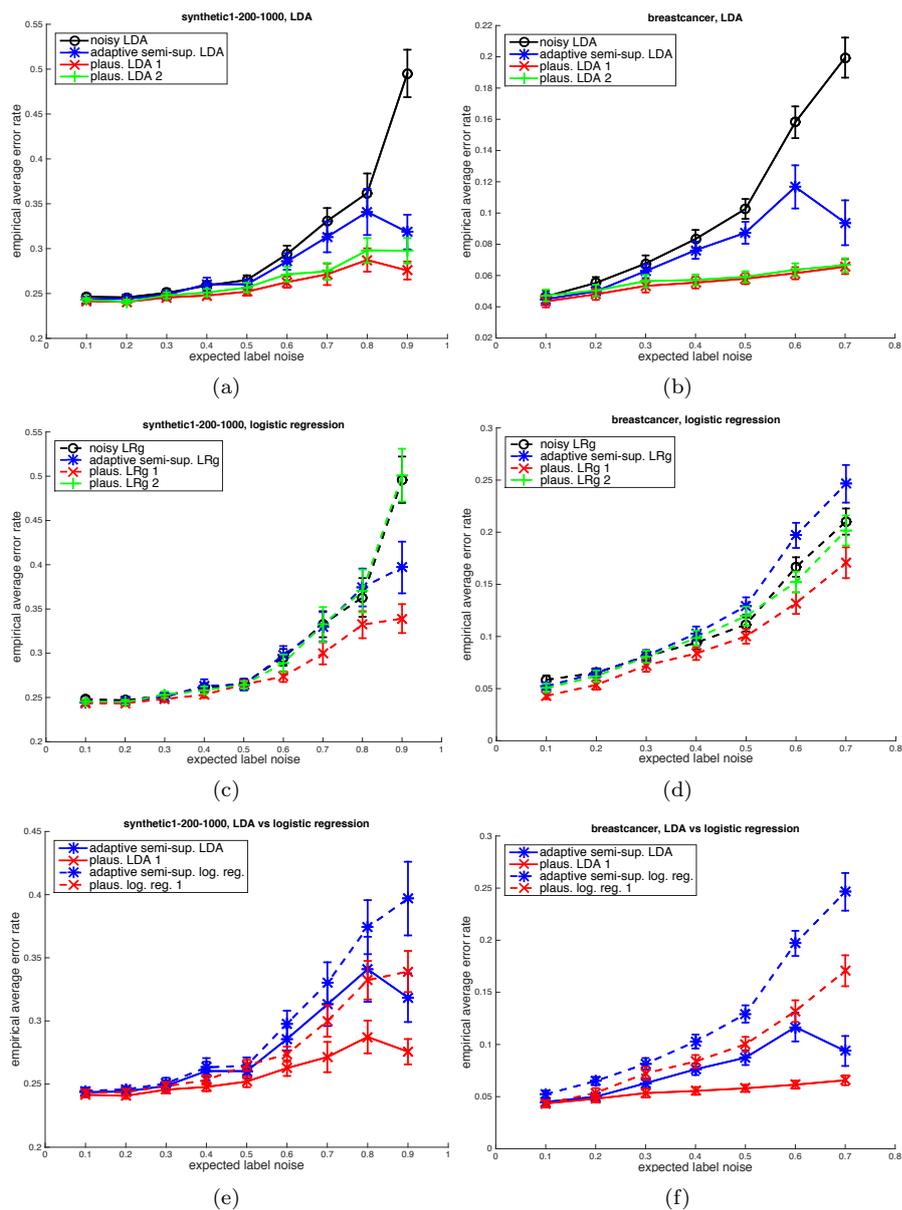


Fig. 1: synthetic (a, c, e) and breastcancer (b, d, f) data. Top: LDA, Middle: LR, Bottom: LDA vs. LR. Black/o: noisy labels, blue/*: adaptive semi-supervised learning, red/x: soft labels based on individual noise level p_i , green/+ : soft labels based on mean noise level \bar{p} ; plain lines: LDA, dashed lines: logistic regression.

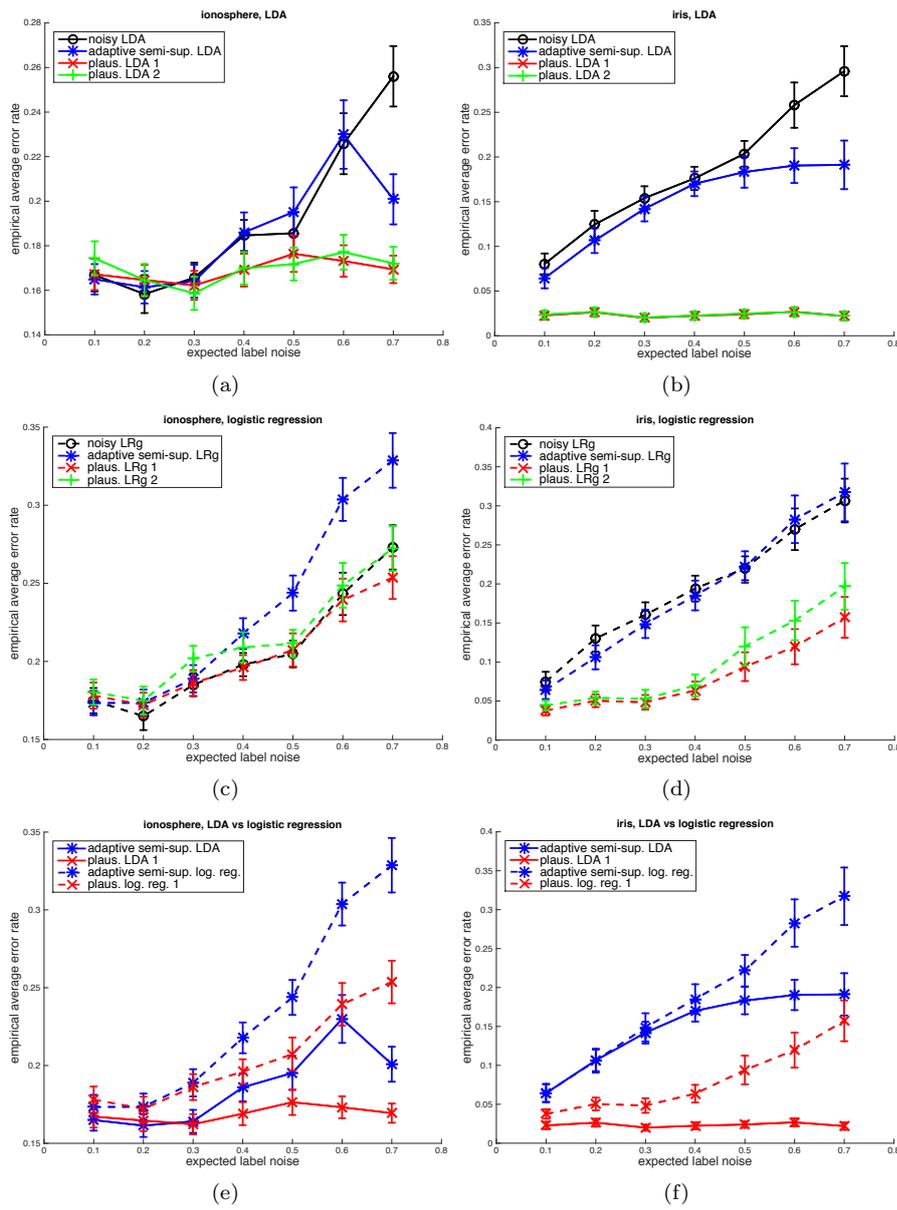


Fig. 2: ionosphere (a, c, e) and iris (b, d, f) data. Top: LDA, Middle: LR, Bottom: LDA vs. LR. Black/o: noisy labels, blue/*: adaptive semi-supervised learning, red/x: soft labels based on individual noise level p_i , green/+ : soft labels based on mean noise level \bar{p} ; plain lines: LDA, dashed lines: logistic regression.

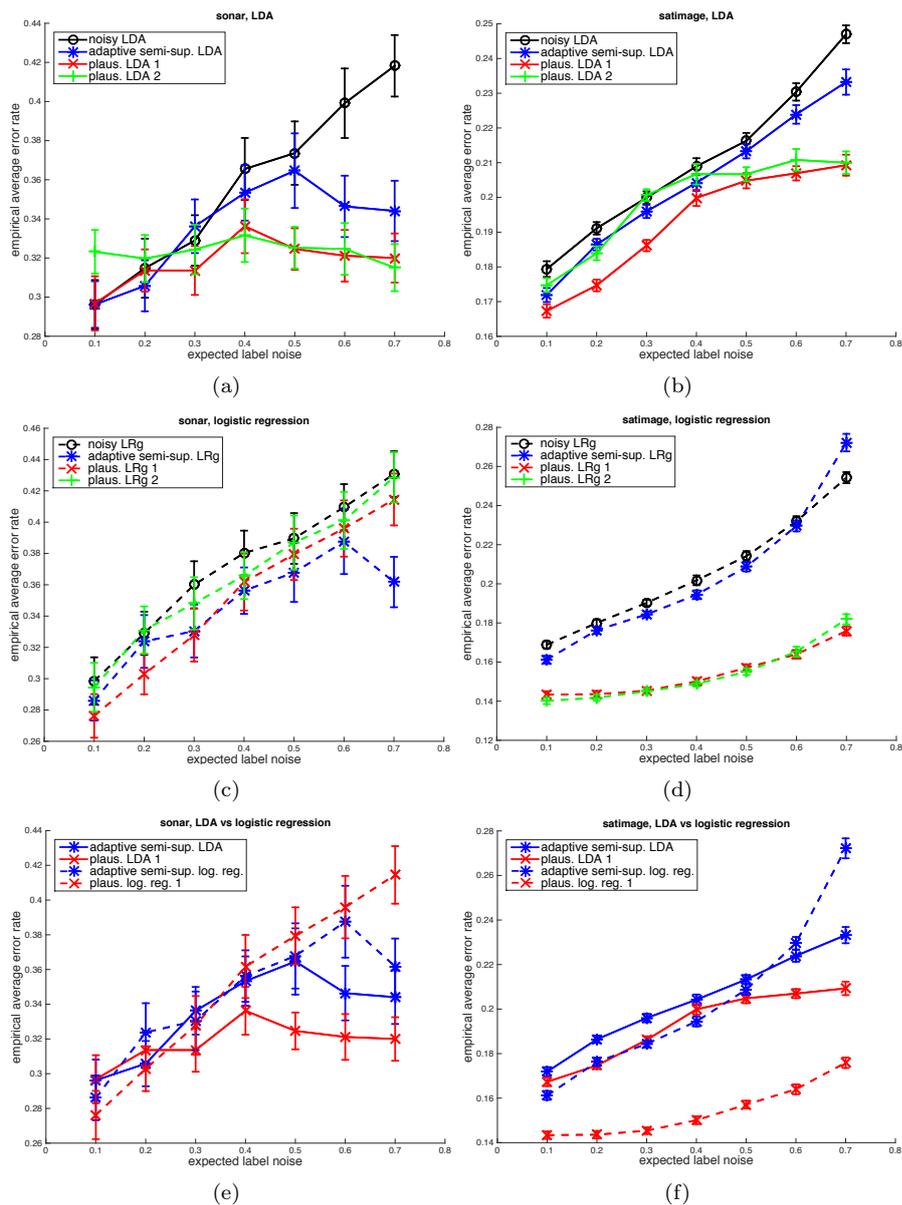


Fig. 3: sonar (a, c, e) and satimage (b, d, f) data. Top: LDA, Middle: LR, Bottom: LDA vs. LR. Black/o: noisy labels, blue/*: adaptive semi-supervised learning, red/x: soft labels based on individual noise level p_i , green/+ : soft labels based on mean noise level \bar{p} ; plain lines: LDA, dashed lines: logistic regression.

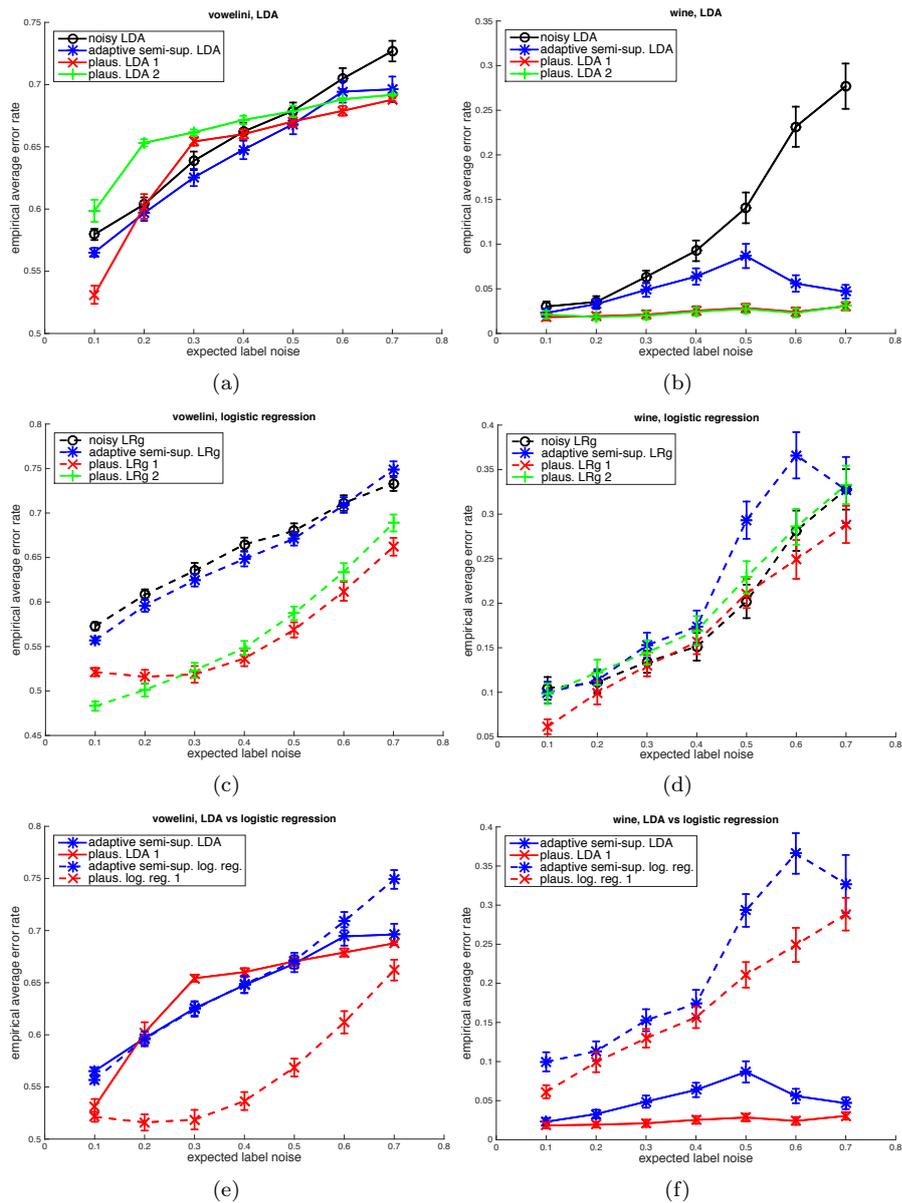


Fig. 4: vowel (a, c, e) and wine (b, d, f) data. Top: LDA, Middle: LR, Bottom: LDA vs. LR. Black/o: noisy labels, blue/*: adaptive semi-supervised learning, red/x: soft labels based on individual noise level p_i , green/+ : soft labels based on mean noise level \bar{p} ; plain lines: LDA, dashed lines: logistic regression.

5.2 Real applications

In this section, we present two real applications where class labels are elicited from human subjects or “experts”. The first application described in Section 5.2.1 concerns the detection of K -complexes in EEG data. In this application, the ground truth cannot be determined objectively and we need to have the data labeled by physicians. Uncertainty is reflected by disagreement between experts. The second application, described in Section 5.2.2, is about expression recognition in facial images. The expression on a face cannot always be determined unambiguously by human subjects. Five subjects were asked to rate the plausibility of each of the six basic expressions for each image, and plausibility assessments were combined by Dempster’s rule. In each of the two applications, we will compare the performances of LDA and those of LR, both with soft labels and with crisp labels ignoring labeling uncertainty.

5.2.1 K -complex detection in EEG sleep data

The methods introduced in this paper were applied to the problem of detecting K -complexes in sleep EEG, using the data described in [31, 32]. The K -complex is a transient EEG pattern which plays a major role in sleep stage assessment. It has a duration of 500 to 1500 ms, and is characterized by a sharp upward wave followed by a downward one. Its amplitude is three times background activity [32]. The discrimination of K -complexes from background activity is generally considered as a very complex pattern recognition problem.

The data used in this experiment consisted of 1178 EEG signals encoded as 64-dimensional patterns. Some of these signals were negative examples containing paroxysmal delta bursts, a phenomenon bearing some resemblance to K -complexes. The other signals consisted of patterns which, after visual inspection by five physicians, had been classified as containing a K -complex by at least one of them. Among these examples, those categorized in the K -complex class by a majority of experts were considered as positive examples, the others as negative ones. Each example (positive or negative) was then assigned a soft label consisting of a Bayesian mass function m_i such that

$$m_i(\{1\}) = k_i/5, \quad m_i(\{0\}) = 1 - k_i/5, \quad (29)$$

where 1 and 0 represent, respectively, the positive (K -complex) and negative (delta wave) class, and k_i denotes the number of experts who classified the pattern as positive. The number of cases for each value of k_i are shown in Table 2. In total, $116 + 88 + 59 = 263$ instances (22.33%) were classified as positive by a majority of experts, while the remaining 915 instances (77.67%) are almost certainly negative, or were not recognized as positive by a majority of experts. Examples of signals with their classification by the five physicians are shown in Figure 5.

Both LR and LDA were applied to these data. To reduce the input dimension, Principal Component Analysis (PCA) was first used as a preprocessing

Table 2: Sleep data: number of instances n_k classified in the positive class by k experts.

k	0	1	2	3	4	5
n_k	397	340	178	116	88	59

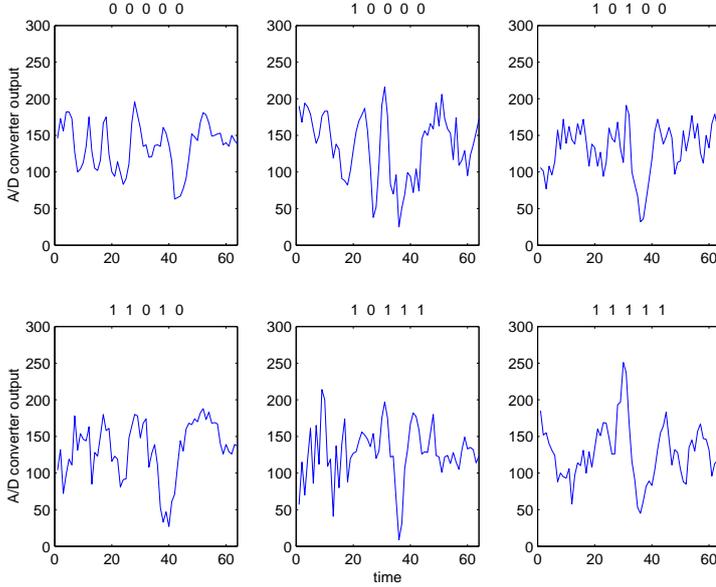


Fig. 5: Six examples of EEG signals. The figures above each plot indicates the classification given of the five experts (1 for K -complex and 0 for delta-wave).

step, and the number of components was varied between 1 and 20. The LR and LDA classifiers were trained using three different sets of labels:

1. Soft labels (29), taking into account the proportion of experts in favor of each class;
2. Crisp labels, corresponding to the majority decision;
3. “Semi-supervised labels”: instances classified as positive by two or three experts were considered as ambiguous and were labeled by the vacuous mass function $m_?$; the other instances were labeled unambiguously according to the majority class.

The results were evaluated using two measures: the expected loss (28), and the error rate, assuming the majority class to be the true class. We note that, as mass functions m_i are Bayesian, the lower and upper expected losses are equal in this case. We used 10-fold cross-validation, repeated 10 times with different random partitions. The mean cross-validation expected loss and error

rate with corresponding 95% confidence intervals are represented as functions of the number of principal components for the LR and LDA with the three sets of labels in Figure 6. For LR, training the classifier with soft labels clearly improves the performances, according to both criteria (Figures 6(a) and 6(b)). For LDA, this is true for 5 to 8 principal components, with non significant differences for smaller or larger numbers of components (Figures 6(c) and 6(d)). For both methods, the best results are obtained with 12 principal components. Figure 7 shows boxplots of expected losses and error rates for LR and LDA with 12 principal components, trained using with the three label sets. Logistic regression outperforms LDA for this dataset, and the best results are obtained when training the classifier with soft labels, whatever the performance measure.

For this dataset, LR was thus clearly able to exploit the information contained in soft labels to construct a better classifier. This phenomenon can be explained by the fact that ambiguous patterns, which have a great chance of being mislabeled, have less informative soft labels. Their influence on the final parameter estimates is thus downweighted by the evidential EM algorithm.

5.2.2 Facial expression recognition

As another application of partially supervised learning with soft labels, we considered the task of recognizing facial expressions from face images. Here, soft labels may arise from the difficulty of labeling faces with a single expression unambiguously. We considered 216 images data from the CMU-Pittsburgh image database [21], with 36 images for each of the six basic expressions (see Figure 8). Following [14], we used aligned and cropped versions of these images of size 60×70 , corresponding to vectors of 4200 pixels.

The database was split into a learning set and a test set of 108 images each (with 18 images per expression). Images in the learning set were viewed by five subjects, who assigned a soft label to each image by assessing plausibilities for each of the six expressions. Figure 9 shows two examples of images with the corresponding plausibility assessments, for which there was some disagreement between assessors. For instance, for the image of Figure 9(a), one subject considered that it expresses disgust for sure (white horizontal bar), while others hesitated between anger and disgust, or between disgust and sadness. The assessments from the five experts were discounted [33] and combined by Dempster's rule using (3). For image i , the unnormalized contour function pl_i^* was thus

$$pl_i^*(k) = \prod_{j=1}^5 [0.1 + 0.9 pl_{ij}(\omega_k)], \quad (30)$$

where $pl_{ij}(k)$ is the plausibility of expression k for image i assessed by subject j . The plausibilities were then normalized to ensure that $\max_k pl_i(k) = 1$. Figure 10 shows two examples of individual and combined plausibility assessments. We can remark that combination rule (30) eliminates alternatives considered as impossible by most subjects, even if they were considered as

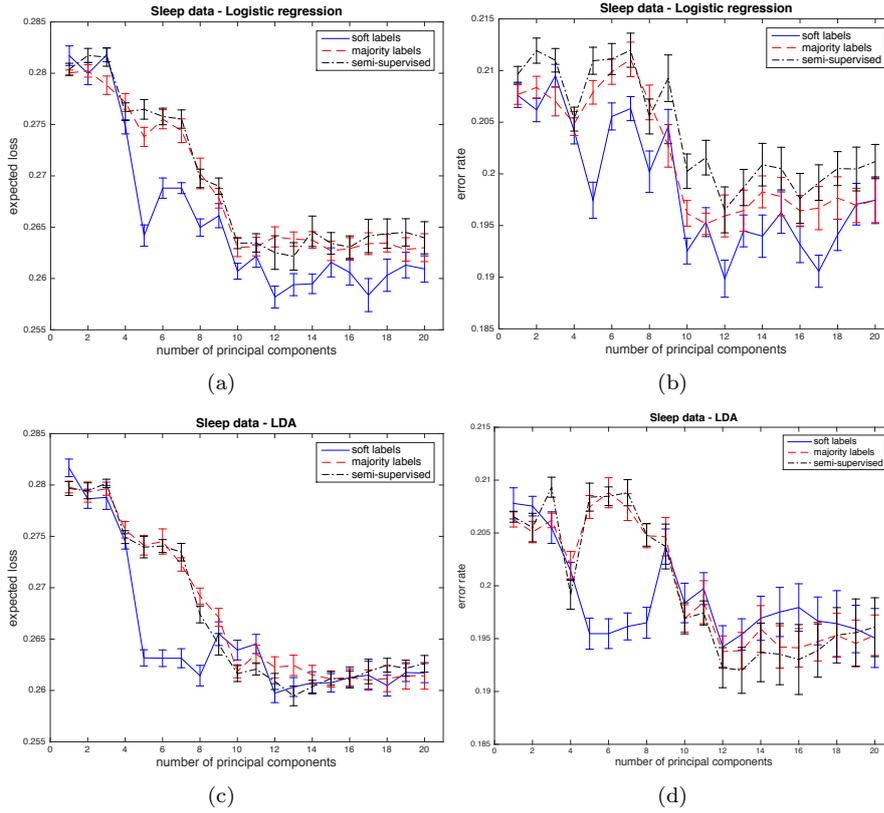


Fig. 6: Performances of LR (a, b) and LDA (c, d) for the sleep data, as a function of the number of principal components. The criteria are the expected loss (a,c) and the error rate taking the majority labels as ground truth (b, d). In each graph, the three curves corresponds to three sets of labels: soft labels, majority labels, and “semi-supervised” labels (see details in text).

completely plausible by one assessor, such as anger in Figure 10(a) or sadness in Figure 10(b).

As in Section 5.2.1, PCA was used to reduce the dimension of the data. Logistic regression and LDA were then applied to the data with different numbers of principal components, and two sets of labels: the combined soft labels, and crisp labels corresponding to the most plausible expression. For logistic regression, we used quadratic penalization with coefficient $\lambda = 10^{-4}$. The results are reported in Figure 11, which shows the test error rate as a function of the number of components, for each of the two classifiers and the two sets of labels. For this dataset, LDA clearly outperforms LR, with the lowest error rate obtained for around 30 principal components. For LDA, the best results are obtained with the soft labels, which can be explained by the smaller im-

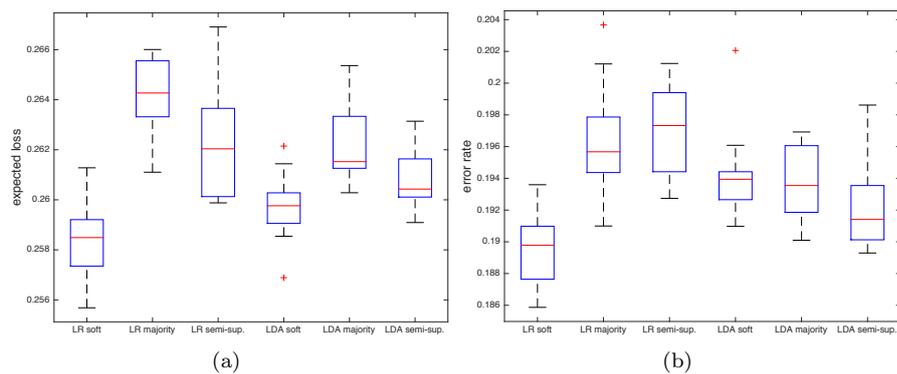


Fig. 7: Expected losses (a) and error rates (b) for classifiers trained with 12 principal components as input (sleep data). The classifiers are, from left to right: LR with soft labels, majority labels and semi-supervised labels; LDA with soft labels, majority labels and semi-supervised labels.

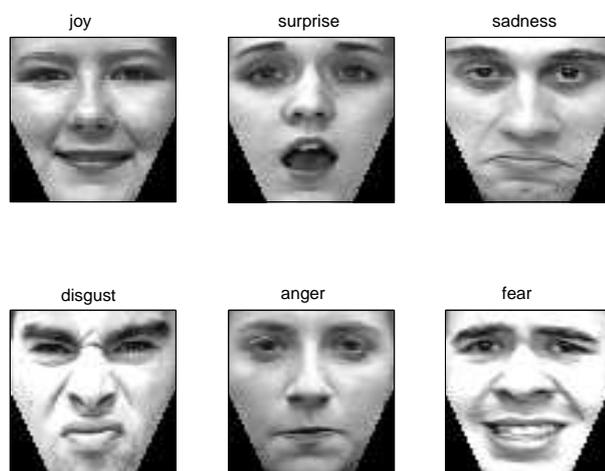


Fig. 8: Examples of each of the six basic expressions.

portance given to ambiguous cases. However, the LDA classifier trained with soft labels is more prone to overfitting, as shown by the sharp increase of error rate with the input dimension. In contrast, LR does not work too well for this data, with a lowest error rate almost twice as large as that obtained by LDA. For LR, there also seems to be some advantage of using soft labels for input dimensions larger than 40, but this advantage is not so clear as it is for LDA.

This second experiment using a real data set with soft labels provides additional evidence of the potential gain of using soft labels and confirms

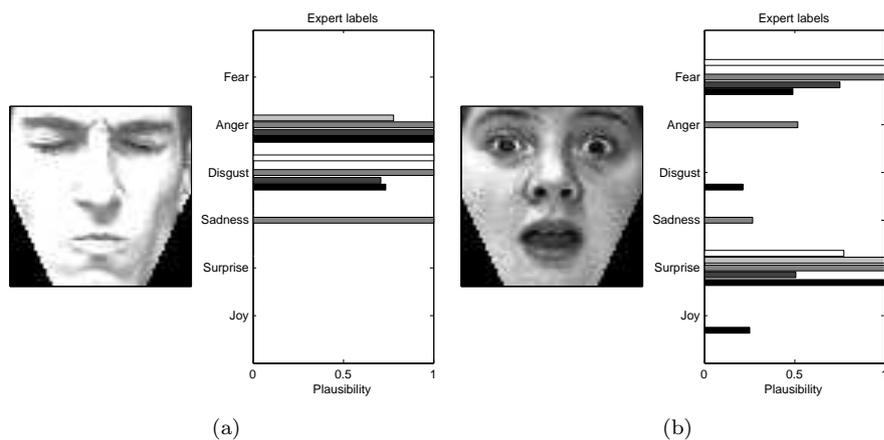


Fig. 9: Two examples of plausibility assessments provided by the five subjects. The plausibility values are shown as horizontal bars next to each of the two images, with one gray level for each subject.

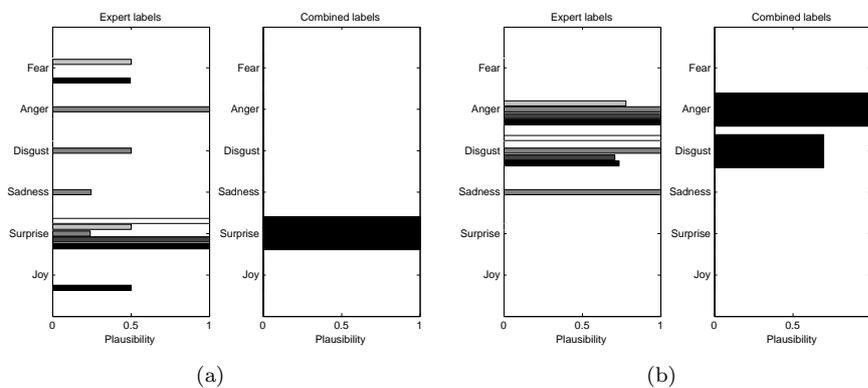


Fig. 10: Two examples of individual and combined plausibility assessments.

similar findings reported in [3] or [30], for instance. As with the sleep data analyzed in Section 5.2.1, this gain of performance can be explained by the smaller influence of potentially mislabeled data on the final classifier.

6 Conclusions

Whereas a lot of methods exist for supervised or unsupervised classification, we have seen in recent years a growing interest for more general forms of learning tasks, in which information about the class of learning instances is partial. In this paper, we have considered a very general framework for address-

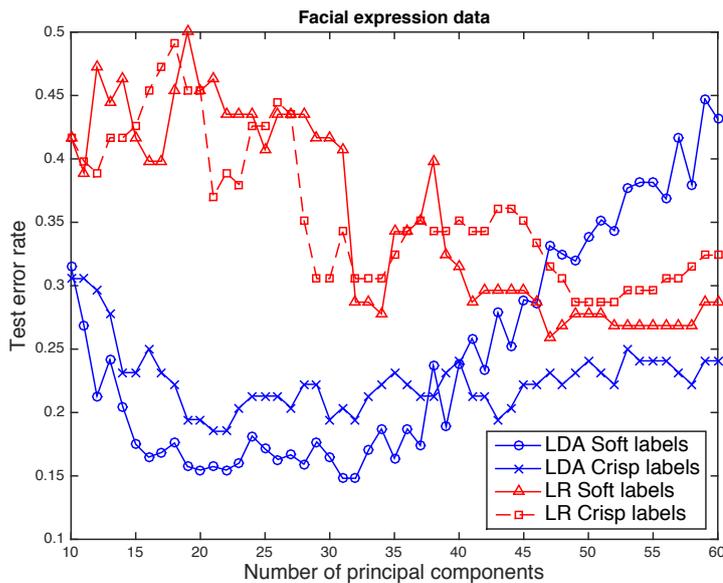


Fig. 11: Facial expression data: test error rates vs. number of principal components for LDA and LR, with soft and crisp labels.

ing such problems, which consists in representing partial class information by Dempster-Shafer mass functions constituting soft labels. Two types of linear classifiers have been considered: LDA based on a generative model, and LR based on a discriminative model. Both classifiers can be trained in a partially supervised mode by maximizing an evidential likelihood function, which can be done efficiently using the evidential EM algorithm.

As a generative model represents the joint distribution of the input vector and the response variable, it can be trained in an unsupervised setting; consequently, it can be expected to perform well with scant class information. In contrast, a discriminative model relies on weaker assumptions and can be expected to perform better than a generative model when the model is inadequate, provided sufficient class information is available. Our experimental results support these assumptions. Using artificial and real data for which class labels have been perturbed by noise, we have shown that soft labels taking into account labelling uncertainty can be successfully exploited to achieve significantly lower error rates, as compared to precise but noisy labels. Both LDA and LR classifiers are able to exploit partial information contained in soft labels. However, LDA usually performs better, especially when the mislabeling probability is high, in which case the soft labels become very imprecise and learning becomes almost unsupervised. However, LR outperforms LDA for some datasets. It is thus necessary to consider several models and select

the best ones, which can also be done in a partially supervised setting using lower and upper expected losses.

We have also considered two real applications in which soft labels naturally arise from the absence of ground truth class information: detection of K -complex in EEG sleep data and recognition of facial expressions in images. In these two cases, soft labels have been shown to allow for better performances, as compared to using the most plausible label for each instance and ignoring labeling uncertainty. Whereas large amounts of unlabeled data (e.g., images) can be retrieved from the internet or can be generated by sensors, reliably labeled data remain relatively scarce because of the cost of visually inspecting the data to determine their class, when no ground truth is available. In this context, our results suggest that partial class information can be useful, provided labeling uncertainty is suitably represented in the form of soft labels.

Although linear classifiers have been exclusively considered in this study for simplicity, there is obviously no conceptual difficulty with extending our approach to non linear parametric classifiers such as, e.g., quadratic or mixture discriminant analysis, generalized additive models, kernel logistic regression, etc. More fundamentally, other likelihood-based approaches to inference from uncertain data, such as proposed in [18] or [6], could be considered and compared with our method. Finally, the elicitation of soft labels from single or multiple experts is also an important topic that remains to be thoroughly investigated.

Acknowledgements The author thanks Cédric Richard and Régis Lengellé for providing the EEG data, as well as the Foundation for Applied Neuroscience Research in Psychiatry (CHS de Rouffach, 68250 Rouffach, France) for authorizing their use. They also thank Franck Davoine for providing the facial expression data. This research was supported by a grant from the Beijing Government as part of the Overseas Talents program.

References

1. Bishop CM (2006) Pattern Recognition and Machine Learning. Springer
2. Chapelle O, Schölkopf B, Zien A (eds) (2006) Semi-Supervised Learning. MIT Press, Cambridge, Ma
3. Cherfi ZL, Oukhellou L, Côme E, Dencœux T, Aknin P (2012) Partially supervised independent factor analysis using soft labels elicited from multiple experts: Application to railway track circuit diagnosis. *Soft Computing* 16(5):741–754
4. Côme E, Oukhellou L, Dencœux T, Aknin P (2009) Learning from partially supervised data using mixture models and belief functions. *Pattern Recognition* 42(3):334–348
5. Cour T, Sapp B, Taskar B (2011) Learning from partial labels. *Journal of Machine Learning Research* 12:1225–1261
6. Couso I, Dubois D (2017) Maximum likelihood under incomplete information: Toward a comparison of criteria. In: Ferraro MB, Giordani P,

- Vantaggi B, Gagolewski M, Ángeles Gil M, Grzegorzewski P, Hryniewicz O (eds) *Soft Methods for Data Science*, Springer International Publishing, Cham, pp 141–148
7. Dempster AP (1967) Upper and lower probabilities induced by a multi-valued mapping. *Annals of Mathematical Statistics* 38:325–339
 8. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* 39:1–38
 9. Denceux T (1995) A k -nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans on Systems, Man and Cybernetics* 25(05):804–813
 10. Denceux T (2013) Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Transactions on Knowledge and Data Engineering* 25(1):119–130
 11. Denceux T (2014) Likelihood-based belief function: justification and some extensions to low-quality data. *International Journal of Approximate Reasoning* 55(7):1535–1547
 12. Denceux T, Skarstein-Bjanger M (2000) Induction of decision trees for partially classified data. In: *Proceedings of SMC’2000*, IEEE, Nashville, TN, pp 2923–2928
 13. Denceux T, Zouhal LM (2001) Handling possibilistic labels in pattern classification using evidential reasoning. *Fuzzy Sets and Systems* 122(3):47–62
 14. Dubuisson S, Davoine F, Masson MH (2002) A solution for facial expression representation and recognition. *Signal Processing: Image Communication* 17(9):657 – 673
 15. Elouedi Z, Mellouli K, Smets P (2001) Belief decision trees: Theoretical foundations. *International Journal of Approximate Reasoning* 28:91–124
 16. Hasan A, Wang Z, Mahani A (2016) Fast estimation of multinomial logit models: R package `mnlogit`. *Journal of Statistical Software* 75(1):1–24
 17. Heitjan DF, Rubin DB (1991) Ignorability and coarse data. *Ann Statist* 19(4):2244–2253
 18. Hullermeier E (2014) Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning* 55(7):1519 – 1534
 19. Hullermeier E, Beringer J (2005) Learning from ambiguously labeled examples. In: *Proceedings of the 6th International Symposium on Intelligent Data Analysis (IDA-05)*, Madrid, Spain
 20. Jaffray JY (1989) Linear utility theory for belief functions. *Operations Research Letters* 8(2):107 – 112
 21. Kanade T, Cohn J, Tian Y (2000) Comprehensive database for facial expression analysis. In: *Proceedings of the Fourth International Conference of Face and Gesture Recognition*, Grenoble, France, pp 46–53
 22. Li J (2013) Logistic regression. Course Notes, <http://sites.stat.psu.edu/~jiali/course/stat597e/notes2/logit.pdf>
 23. Ma L, Destercke S, Wang Y (2016) Online active learning of decision trees with evidential data. *Pattern Recognition* 52:33–45

24. McLachlan GJ, Krishnan T (1997) *The EM Algorithm and Extensions*. Wiley, New York
25. McLachlan GJ, Peel D (2000) *Finite Mixture Models*. Wiley, New York
26. Nguyen N, Caruana R (2008) Classification with partial labels. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, NY, USA, KDD '08, pp 551–559
27. Press SJ, Wilson S (1978) Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association* 73(364):699–705
28. Quost B (2014) Logistic regression of soft labeled instances via the evidential EM algorithm. In: Cuzzolin F (ed) *Belief Functions: Theory and Applications: Third International Conference, BELIEF 2014*, Oxford, UK, September 26-28, 2014. *Proceedings*, Springer International Publishing, Cham, pp 77–86
29. Quost B, Denœux T (2016) Clustering and classification of fuzzy data using the fuzzy EM algorithm. *Fuzzy Sets and Systems* 286:134 – 156
30. Ramasso E, Denœux T (2013) Making use of partial knowledge about hidden states in HMMs: an approach based on belief functions. *IEEE Transactions on Fuzzy Systems* 21(6):1–11
31. Richard C (1998) *Une méthodologie pour la détection à structure imposée. applications au plan temps-fréquence*. PhD thesis, Université de Technologie de Compiègne
32. Richard C, Lengellé R (1999) Data driven design and complexity control of time-frequency detectors. *Signal Processing* 77:37–48
33. Shafer G (1976) *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J.
34. Strat TM (1990) Decision analysis using belief functions. *International Journal of Approximate Reasoning* 4(5–6):391–417
35. Trabelsi S, Elouedi Z, Mellouli K (2007) Pruning belief decision tree methods in averaging and conjunctive approaches. *International Journal of Approximate Reasoning* 46(3):568–595
36. Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems* 1:3–28