Logistic regression revisited: belief function analysis

Thierry Denoeux

Université de Technologie de Compiègne, CNRS,
UMR 7253 Heudiasyc, France
tdenoeux@utc.fr

Abstract. We show that the weighted sum and softmax operations performed in logistic regression classifiers can be interpreted in terms of evidence aggregation using Dempster's rule of combination. From that perspective, the output probabilities from such classifiers can be seen as normalized plausibilities, for some mass functions that can be laid bare. This finding suggests that the theory of belief functions is a more general framework for classifier construction than is usually considered.

Keywords: Evidence theory, Dempster-Shafer theory, classification, machine learning.

1 Introduction

In the last twenty years, the Dempster-Shafer (DS) theory of belief functions has been increasingly applied to 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., [8], [1], [7]). Another approach is to design *evidential classifiers*, which can be defined as classifiers built from basic principles of DS theory. Typically, an evidential classifier has the structure depicted in Figure 1: when presented by a feature vector x, the system computes k mass functions $m_1, \ldots, m_k$ defined on the set Θ of classes, based on a learning set. These mass functions are then combined using Dempster's rule, or any other rule. The first evidential classifier was the evidential k-nearest neighbor classifier [3], in which mass functions m_j are constructed from the k nearest neighbor of x, and combined by Dempster's rule. In the evidential neural network classifier [5], a similar principle is applied, but 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 distance-based classifiers, but also a broad class of widely-used classifiers, including logistic regression and its nonlinear extensions, can be seen as evidential classifiers. This finding leads us to the conclusion that DS theory is a much more general framework for classifier construction than was initially believed.

The rest of the paper is organized as follows. Some background definitions will first be recalled in Section 2. A general model of feature-based evidence will

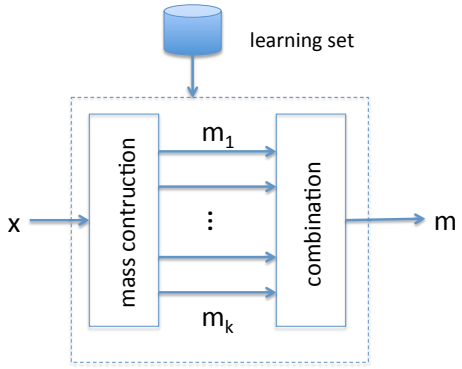


Fig. 1. Basic structure of an evidential classifier.

be described in Section 3, where we will show that the normalized plausibility function, after combining the evidence of J features, is identical to the output of logistic regression. The recovery of the full mass function will then be addressed, and a simple example will be given in Section 4. Section 5 will conclude the paper.

2 Background

In this section, we first recall some basic notions and definitions needed in the rest of the paper. The notion of weight of evidence will first be recalled in Section 2.1, and some notations for logistic regression will be introduced in Section 2.2.

2.1 Weights of evidence

Let us consider a simple mass function m on a frame Θ, such that

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

where s is a degree of support in $[0, 1]$. Typically, such a mass function represents some elementary piece of evidence supporting hypothesis A. Shafer [9, page 77] defines the *weight* of this evidence as $w = -\ln(1 - s)$. Conversely, we thus have $s = 1 - \exp(-w)$. The rationale for this definition is that weights of evidence are additive: if m_1 and m_2 are two simple mass functions focussed on the same subset A, with weights w_1 and w_2, then the orthogonal sum $m_1 \oplus m_2$ corresponds to the weight $w_1 + w_2$. If we denote a simple mass function with focal set A and weight w by A^w, we thus have $A^{w_1} \oplus A^{w_2} = A^{w_1+w_2}$. It follows that any separable mass function can be written as $m = \bigoplus_{\emptyset \neq A \subset \Theta} A^{w_A}$, where w_A is the weight of evidence pointing to A. We note that, in [6], following [10], 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.

2.2 Logistic regression

Consider a multi-category[1] classification problem with J-dimensional feature vector $x = (x_1, \ldots, x_J)$ and class variable $Y \in \Theta = \{\theta_1, \ldots, \theta_K\}$ with $K > 2$. In the logistic regression model, we assume the logarithms of the posterior class probabilities $\mathbb{P}(Y = \theta_k|x)$ to be affine functions of x, i.e.,

$$\ln \mathbb{P}(Y = \theta_k|x) = \sum_{j=1}^{J} \beta_{jk} x_j + \beta_{0k} + \gamma, \quad \forall k \in [\![1, K]\!], \tag{1}$$

where β_{jk}, $j = 0, \ldots, J$ are parameters and γ is a constant. Using the equation $\sum_{k=1}^{K} \mathbb{P}(Y = \theta_k|x) = 1$, we easily get the following expressions for the posterior probabilities,

$$\mathbb{P}(Y = \theta_k|x) = \frac{\exp\left(\sum_{j=1}^{J} \beta_{jk} x_j + \beta_{0k}\right)}{\sum_{l=1}^{K} \exp\left(\sum_{j=1}^{J} \beta_{jl} x_j + \beta_{0l}\right)}. \tag{2}$$

This transformation from arbitrary real quantities (1) to probabilities is sometimes referred to as the *softmax transformation*. Parameters β_{jk} are usually estimated by maximizing the conditional likelihood. In feedforward neural networks with a softmax output layer, a similar approach is used, with variables x_j defined as the outputs of the last hidden layer of neurons. These variables are themselves defined as complex nonlinear functions of the input variables, which are optimized together with the decision layer weights β_{jk}. Logistic regression is functionally equivalent to a feedforward neural network with no hidden layer.

3 Model

We consider a multi-category classification problem as described in Section 2.2. We assume that each feature x_j provides some evidence about the class variable Y. For each θ_k, the evidence of feature x_j points either to the singleton $\{\theta_k\}$ or to its complement $\overline{\{\theta_k\}}$, depending on the sign of

$$w_{jk} = \beta_{jk} x_j + \alpha_{jk}, \tag{3}$$

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, respectively,

$$w_{jk}^{+} = (w_{jk})_{+} \quad \text{and} \quad w_{jk}^{-} = (w_{jk})_{-}, \tag{4}$$

where $(\cdot)_{+}$ and $(\cdot)_{-}$ denote, respectively, the positive and the negative parts. For each feature x_j and each class θ_k, we thus have two simple mass functions

$$m_{jk}^{+} = \{\theta_k\}^{w_{jk}^{+}} \quad \text{and} \quad m_{jk}^{-} = \overline{\{\theta_k\}}^{w_{jk}^{-}}. \tag{5}$$

[1] The case of binary classification with $K = 2$ classes requires a separate treatment. Due to space constaints, we focus on the multi-category case in this paper.

Assuming these mass functions to be independent, they can be combined by Dempster's rule. Let

$$m_k^+ = \bigoplus_{j=1}^{J} m_{jk}^+ = \{\theta_k\}^{w_k^+} \text{ and } m_k^- = \bigoplus_{j=1}^{J} m_{jk}^- = \overline{\{\theta_k\}}^{w_k^-}$$

where

$$w_k^+ = \sum_{j=1}^{J} w_{jk}^+ \quad \text{and} \quad w_k^- = \sum_{j=1}^{J} w_{jk}^-. \tag{6}$$

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

$$m^+ = \bigoplus_{k=1}^{K} m_k^+ \quad \text{and} \quad m^- = \bigoplus_{k=1}^{K} m_k^-,$$

and let pl^+ and pl^- be the corresponding contour functions. We have

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

and

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

Finally, let $m = m^+ \oplus m^-$ and let pl be the corresponding contour function. We have

$$\begin{aligned} pl(\theta_k) &\propto pl^+(\theta_k) pl^-(\theta_k) \propto \exp(w_k^+ - w_k^-) \\ &\propto \exp\left(\sum_{j=1}^{J} w_{jk}\right) = \exp\left(\sum_{j=1}^{J} \beta_{jk} x_j + \sum_{j=1}^{J} \alpha_{jk}\right). \end{aligned}$$

Let p be the probability mass function induced from m by the plausibility-probability transformation [2], and let

$$\beta_{0k} = \sum_{j=1}^{J} \alpha_{jk}. \tag{7}$$

We have

$$p(\theta_k) = \frac{\exp\left(\sum_{j=1}^{J} \beta_{jk} x_j + \beta_{0k}\right)}{\sum_{l=1}^{K} \exp\left(\sum_{j=1}^{J} \beta_{jl} x_j + \beta_{0l}\right)}, \tag{8}$$

which is equivalent to (2). We thus have proved that the output probabilities computed by a logistic regression classifier can be seen as the normalized plausibilities obtained after combining elementary mass functions (5) by Dempster's rule: these classifiers are, thus, evidential classifiers as defined in Section 1.

4 Recovering the mass function

Having shown that the output probabilities of logistic regression classifiers are normalized plausibilities, it is interesting to recover the underlying output mass function, defined as

$$m = \bigoplus_{k=1}^{K} \left(\{\theta_k\}^{w_k^+} \oplus \overline{\{\theta_k\}}^{w_k^-} \right). \tag{9}$$

Its complete expression can be derived (after some tedious calculation), but it cannot be given here for lack of space.

There is, however, a difficulty related to the identifiability of the weights w_k^+ and w_k^-. First, parameters β_{jk} are not themselves identifiable, because adding any constant vector $\boldsymbol{c}$ to each vector $\beta_k = (\beta_{0k}, \ldots, \beta_{Jk})$ produces the same normalized plausibilities (8). Secondly, for given β_{0k}, any α_{jk} verifying (7) will yield the same probabilities (8). This problem is addressed in the next section.

4.1 Identification

To identifying the underlying output mass function, we propose to apply the Least Commitment Principle, by searching for the mass function m^* of the form (9) verifying (8) and such that the sum of the squared weights of evidence is minimum. More precisely, let $\{(x_i, y_i)\}_{i=1}^n$ be a learning set, let $\widehat{\beta}_{jk}$ be the maximum likelihood estimates of the weights β_{jk}, and let $\boldsymbol{\alpha}$ denote the vector of parameters α_{jk}. Any $\beta_{jk}^* = \widehat{\beta}_{jk} + c_j$ will verify (8). The parameter values β_{jk}^* and α_{jk}^* minimizing the sum of the squared weights of evidence can thus be found by solving the following minimization problem

$$\min f(\boldsymbol{c}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K} \left[(\widehat{\beta}_{jk} + c_j) x_{ij} + \alpha_{jk} \right]^2 \tag{10}$$

subject to

$$\sum_{j=1}^{J} \alpha_{jk} = \widehat{\beta}_{0k} + c_0, \quad \forall k \in [\![1, K]\!]. \tag{11}$$

In (10), x_{ij} denotes the value of feature j for learning vector x_i. Developing the square in (10), we get

$$f(\boldsymbol{c},\boldsymbol{\alpha}) = \sum_{j,k}(\widehat{\beta}_{jk}+c_j)^2\left(\sum_i x_{ij}^2\right)+n\sum_{j,k}\alpha_{jk}^2+2\sum_{j,k}(\widehat{\beta}_{jk}+c_j)\alpha_{jk}\sum_i x_{ij}. \quad (12)$$

Assuming that the input variables x_j have been centered, we have $\sum_i x_{ij} = 0$ and $\sum_i x_{ij}^2 = s_j^2$, where s_j^2 is the empirical variance of feature x_j. Eq. (12) then simplifies to

$$f(\boldsymbol{c},\boldsymbol{\alpha}) = \sum_{j,k} s_j^2(\widehat{\beta}_{jk}+c_j)^2 + n\sum_{j,k}\alpha_{jk}^2. \quad (13)$$

Due to constraint (11), for any c_0, the second term in the right-hand side of (13) is minimized for $\alpha_{jk} = \frac{1}{J}(\widehat{\beta}_{0k}+c_0)$, for all $j \in [\![1,J]\!]$ and $k \in [\![1,K]\!]$. Hence, the problem becomes

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

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

$$c_j^* = -\frac{1}{K}\sum_{k=1}^{K}\widehat{\beta}_{jk}, \quad \forall j \in [\![0,J]\!]$$

The optimum coefficients are, thus,

$$\beta_{jk}^* = \widehat{\beta}_{jk} - \frac{1}{K}\sum_{l=1}^{K}\widehat{\beta}_{jl}, \quad \forall j \in [\![0,J]\!], \forall k \in [\![1,K]\!]$$

and

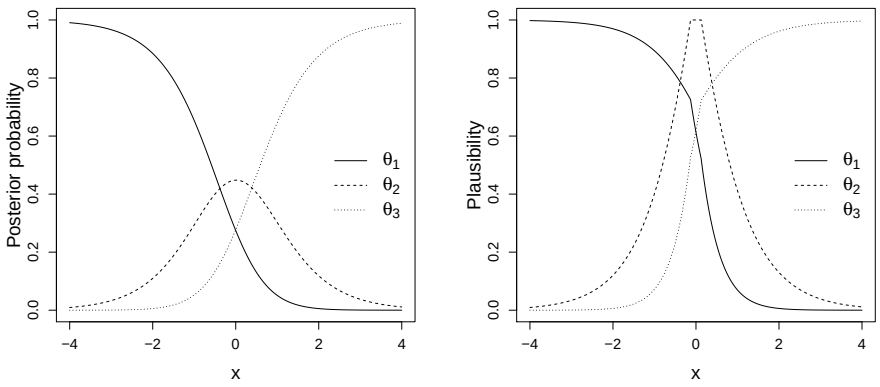
$$\alpha_{jk}^* = \beta_{0k}^*/J, \quad \forall j \in [\![1,J]\!], \forall k \in [\![1,K]\!]. \quad (14)$$

To get the least committed mass function m^* with minimum sum of squared weights of evidence and verifying (8), we thus need to center the rows of the $(J+1)\times K$ matrix $B = (\beta_{jk})$, set α_{jk}^* according to (14), and compute the weights of evidence w_k^- and w_k^+ from (3), (4) and (6).

4.2 Example

As a simple example, let us consider simulated data with $J = 1$ feature, $K = 3$ classes, and Gaussian conditional distributions $X|\theta_k \sim \mathcal{N}(\mu_k, 1)$, with $\mu_1 = -1$, $\mu_2 = 0$ and $\mu_3 = 1$. We randomly generated $10,000$ from each of the three conditional distributions, we standardized the data and we trained a logistic regression classifier on these data. Decisions are usually based on the posterior

class probabilities $\mathbb{P}(\theta_k|x)$ displayed in Figure 2(a). Figure 3 shows the underlying masses, computed as explained in Section 4.1. As we can see, masses are assigned to subsets of classes in regions where these classes overlap, as could be expected. Figure 2(b) shows the contour functions $pl(\theta_k|x)$ vs x. Interestingly, the graphs of these functions have quite different shapes, as compared to those of the posterior probabilities shown in Figure 2(a). Whereas decisions with probabilistic classifiers are classically based on minimum expected loss, seeing logistic regression classifiers as evidential classifiers opens the possibility to experiment with other rules such as minimum lower or upper expected loss [4] or interval dominance [11].

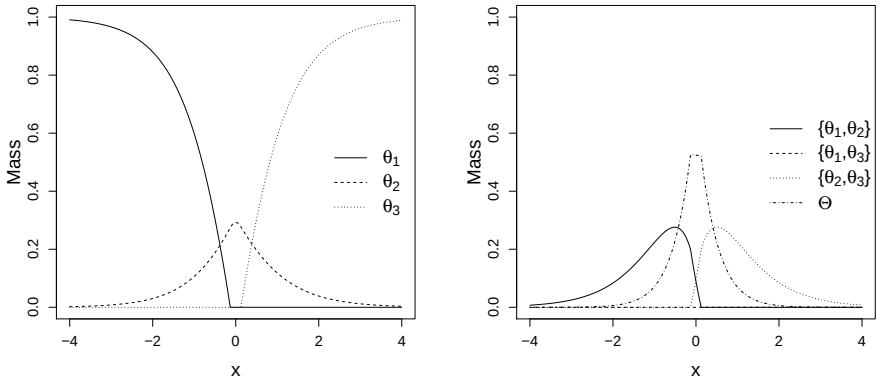


(a) (b)

Fig. 2. Posterior class probabilities $\mathbb{P}(\theta_k|x)$ (a) and contour functions $pl(\theta_k|x)$ for the logistic regression example.

5 Conclusions

We have shown that logistic regression classifiers and also, as a consequence, generalized linear classifiers such as feedforward neural network classifiers, which essentially perform logistic regression in the output layer, can be seen as pooling evidence using Dempster's rule of combination. This finding may have important implications, as it opens the way to a DS analysis of many widely used classifiers, beyond the particular distance-based classifiers introduced in [3] and [5]. In future work, we will deepen this analysis by exploring the consequences of viewing neural network classifiers as evidential classifiers, in terms of decision strategies, classifier fusion, and handling missing or uncertain inputs, among other research directions.



(a) (b)

Fig. 3. Masses on singletons (a) and compound hypotheses (b) vs. x for the logistic regression example.

References

1. Y. Bi, J. Guan, and D. Bell. The combination of multiple classifiers using an evidential reasoning approach. *Artificial Intelligence*, 172(15):1731–1751, 2008.
2. B. R. Cobb and P. P. Shenoy. On the plausibility transformation method for translating belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3):314–330, 2006.
3. T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics*, 25(05):804–813, 1995.
4. T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7):1095–1107, 1997.
5. T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics A*, 30(2):131–150, 2000.
6. T. Denœux. Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. *Artificial Intelligence*, 172:234–264, 2008.
7. B. Quost, M.-H. Masson, and T. Denœux. Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules. *International Journal of Approximate Reasoning*, 52(3):353–374, 2011.
8. G. Rogova. Combining the results of several neural network classifiers. *Neural Networks*, 7(5):777–781, 1994.
9. G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J., 1976.
10. P. Smets. The canonical decomposition of a weighted belief. In *Int. Joint Conf. on Artificial Intelligence*, pages 1896–1901, San Mateo, Ca, 1995. Morgan Kaufman.
11. M. C. Troffaes. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*, 45(1):17 – 29, 2007.