Any Reasonable Cost Function Can Be Used for a Posteriori Probability Approximation

Marco Saerens*, Patrice Latinne †& Christine Decaestecker‡
Université Catholique de Louvain and Université Libre de Bruxelles
Belgium

October 18, 2002

Abstract

In this paper, we provide a straightforward proof of an important, but nevertheless little known, result obtained by Lindley in the framework of subjective probability theory. This result, once interpreted in the machine learning/pattern recognition context, puts new lights on the probabilistic interpretation of the output of a trained classifier. A learning machine, or more generally a model, is usually trained by minimizing a criterion – the expectation of the cost function – measuring the discrepancy between the model output and the desired output. In this letter, we first show that, for the binary classification case, training the model with any 'reasonable cost function' can lead to Bayesian a posteriori probability estimation. Indeed, after having trained the model by minimizing the criterion, there always exists a computable transformation that maps the output of the model to the Bayesian a posteriori probability of the class membership given the input. Then, necessary conditions allowing the computation of the transformation mapping the outputs of the model to the a posteriori probabilities are derived for the multi-output case. Finally, these theoretical results are illustrated through some simulation examples involving various cost functions.

*Marco Saerens is with the Information Systems Research Unit, IAG, Université Catholique de Louvain, 1 Place des Doyens, B-1348 Louvain-la-Neuve, Belgium. Email: saerens@isys.ucl.ac.be

†Patrice Latinne is with the IRIDIA Laboratory (Artificial Intelligence Laboratory), cp 194/6, Université Libre de Bruxelles, 50 avenue Franklin Roosevelt, B-1050 Brussels, Belgium. Email: platinne@ulb.ac.be.

‡Christine Decaestecker is a Research Associate with the Belgian Research Founds (F.N.R.S.) at the Laboratory of Histopathology, cp 620, Université Libre de Bruxelles, 808 route de Lennik, B-1070 Brussels, Belgium. Email: cdecaes@ulb.ac.be.

1. Introduction

An important problem concerns the probabilistic interpretation to be given to the output of a learning machine, or more generally a model, after training. It appears that this probabilistic interpretation depends on the cost function used for training. Classification models are almost always trained by minimizing a given criterion, the expectation of the cost function. It is therefore of fundamental importance to have a precise idea of what can be achieved with the choice of this criterion.

Consequently, there has been considerable interest in analyzing the properties of the mean square error criterion – the most commonly used criterion. It is for instance well-known that artificial neural nets (or more generally any model), when trained using the mean square error criterion, produce as output an approximation of the expected value of the desired output conditional on the explanatory input variables if 'perfect training' is achieved (see for instance [1],[5]). We say that **perfect training** is achieved if

- A minimum of the criterion is indeed reached after training, and
- The learning machine is a 'sufficiently powerful model' that is able to approximate the optimal estimator to any degree of accuracy (perfect model matching property).

It has also been shown that other cost functions, for instance the cross-entropy between the desired output and the model output in the case of pattern classification, lead to the same property of approximating the conditional expectation of the desired output as well. We may therefore wonder what conditions a cost function should satisfy in order that the model output has this property. In 1991, following the results of Hampshire & Pearlmutter [3], Miller, Goodman & Smyth [7], [8] answered to this question by providing conditions on the cost function ensuring that the output of the model approximates the conditional expectation of the desired output given the input, in the case of perfect training. These results were rederived by Saerens by using the calculus of variations [9], and were then extended to the conditional median [10]. Also, in [10], a close relationship between the conditions on the cost function ensuring that the output of the model approximates the conditional probability of the desired output given the input, when the performance criterion is minimized, and the quasi-likelihood functions used in the context of applied statistics (generalized linear models; see [6]) was pointed out.

In this work, we focus on **classification**, in which case the model will be called a classifier. In this framework, we show that, for the binary classification case, training the classifier with **any reasonable cost function** leads to a posteriori probability estimation. Indeed, after having trained the model by minimizing

the criterion, there always exists a **computable transformation** that maps the output of the model to the a posteriori probability of the class label. This means that we are free to choose any reasonable cost function we want, and train the classifier with it. We can always remap the output of the model afterwards to the a posteriori probability, for Bayesian decision making. We will see that this property generalizes to a certain extend to the multi-output case.

This important result was proved by Lindley in 1982, in the context of subjective probability theory [4]. Briefly, Lindley considered the case where a person expresses his uncertainty about an event E, conditional upon an event F, by assigning a number, x (we use Lindley's notations). For example, consider a physician who, after the medical examination of a patient, has to express his uncertainty about the diagnosis of a given disease (E), conditional on the result (F) of the examination. This physician then receives a score $f(x, I_E)$ which is function of x and the truth or falsity of E when F is true (where I_E is an indicator variables, i.e. $I_E = 1$ (0) if the event E is true (false)). The score function $f(x, I_E)$ can be interpreted as assigning a penalty or reward depending of the discrepancy between the person's response and the true state of the event E. It is assumed that the person wishes to reduce his expected score. Under a number of reasonable assumptions on the score function f and the possible values x which can be chosen by the person, Linsley proved that there exists a simple transform of the values x which map them on probabilities. This transform is a function of the values $f'(x, 0)$ and $f'(x, 1)$, the differentials of f.

In the present paper, we show that Lindley's approach can be applied in the machine learning/pattern recognition context in the case of pattern classification problems, leading to an interesting result concerning the cost functions used to train a classifier. Lindley's derivation was based on geometrical facts and reasoning, while our proof relies on standard differential calculus, and partly extends to the multiple class problem.

In the following sections, we first introduce the binary output problem from an estimation theory perspective (section 2). Then, we derive the transformation that must be applied to the output of the model to obtain the a posteriori probability of the desired output, given the input (section 3). Some results for the multi-output case are provided in section 4. Finally, examples of cost functions and corresponding mappings to a posteriori probabilities are presented in section 5. We conclude in section 6.

2. Statement of the two-class problem

Let us consider that we are given a sequence of N independent m-dimensional training patterns $\mathbf{x}_k = [x_1(k), x_2(k), \ldots, x_m(k)]^{\mathrm{T}}$, with $k = 1, 2, \ldots, N$, as well as corresponding scalar desired outputs $y_k \in \{0, 1\}$ providing information about the class label of the pattern. If the observation $\mathbf{x}_k$ is assigned to the class label ω_0,

then $y_k = 0$; if it is assigned to the class label ω_1, then $y_k = 1$. The $\mathbf{x}_k$ and the y_k are realizations of the random variables $\mathbf{x}$ and y. We hope that the random vector $\mathbf{x}$ provides some useful information that allows to predict the class label y with a certain accuracy. The objective is to train a model, say a neural network, in order to supply outputs $\widehat{y}_k$ (we assume $0 \leq \widehat{y}_k \leq 1$) that are 'accurate' (in some predefined manner; see below) estimations or predictions of the desired outputs y_k:

$$\widehat{y}_k = N[\mathbf{x}_k, \mathbf{w}] \text{ with } 0 \leq \widehat{y}_k \leq 1 \tag{2.1}$$

where $N[.,.]$ is the function provided by the model, $\mathbf{x}_k$ the input vector (the vector of explanatory variables) supplied to the model, and $\mathbf{w}$ is the parameter vector of the model. In order to measure how 'accurate' is the estimation (2.1), we define a **cost function** (or loss function, penalty function, objective function, empirical risk measure, scoring rule) that provides us a measure of the discrepancy between the **predicted value** $\widehat{y}_k$ and the **desired value** y_k: $\pounds[\widehat{y}_k; y_k]$. The purpose of the training is, of course, to estimate the parameters that minimize this cost.

Since it is generally not possible to minimize the cost function for each k because of the presence of noise or disturbances (for a given value of the input $\mathbf{x}$, the desired output is distributed with a probability density function $p(y|\mathbf{x})$), the best we can do is to minimize this cost 'on average'. This leads to the definition of the **performance criterion** $C[\widehat{y}]$:

$$C[\widehat{y}] = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \pounds[\widehat{y}_k; y_k] \tag{2.2}$$

$$= \iint \pounds[\widehat{y}; y]\, p(\mathbf{x}, y)\, d\mathbf{x}\, dy = E_{\mathbf{x}y}\{\pounds[\widehat{y}; y]\} \tag{2.3}$$

where the integral is defined on the Euclidean space $\Re^m \times \Re^1$ and we assume that there are enough samples so that we can rely on the asymptotic form of the performance criterion. $E_{\mathbf{x}y}\{.\}$ is defined as the standard expectation.

It is convenient to rewrite (2.3)

$$C[\widehat{y}] = \int \left\{ \int \pounds[\widehat{y}; y]\, p(y|\mathbf{x})\, dy \right\} p(\mathbf{x})\, d\mathbf{x} \tag{2.4}$$

If we minimize the inner integral of (2.4) for every possible value of $\mathbf{x}$, then $C[\widehat{y}]$ will also be minimized, since $p(\mathbf{x})$ is non negative. We therefore select $\widehat{y}(\mathbf{x})$ in order to minimize the **conditional criterion**

$$C[\widehat{y}|\mathbf{x}] = \int \pounds[\widehat{y}; y]\, p(y|\mathbf{x})\, dy = E_y\{\pounds[\widehat{y}; y]|\mathbf{x}\} \tag{2.5}$$

for every $\mathbf{x}$, where $C[\widehat{y}|\mathbf{x}]$ is a function of both $\widehat{y}$ and $\mathbf{x}$, and $E_y\{.|\mathbf{x}\}$ is the conditional expectation, given $\mathbf{x}$. This means that the minimization of (2.5) can

be performed independently for every $\mathbf{x}$. Moreover, since $\widehat{y}$ is chosen in order to minimize (2.5) for every value of $\mathbf{x}$, this $\widehat{y}$ will be a function of $\mathbf{x}$. The function of $\mathbf{x}$ that minimizes (2.5) will be called the best or **optimal** estimator, and will be denoted by $\widehat{y}^*(\mathbf{x})$.

We assume that this optimal estimator can be approximated to any degree of accuracy by the model, $\widehat{y} = N[\mathbf{x}, \mathbf{w}]$, for some optimal value of the parameters $\mathbf{w} = \mathbf{w}^*$ (perfect parameters tuning: $\widehat{y}^*(\mathbf{x}) = N[\mathbf{x}, \mathbf{w}^*]$). In other words, we are making a 'perfect model matching' assumption. In the Miller, Goodman & Smyth terminology [7], [8], such a model is called a **sufficiently powerful model** that is able to produce the optimal estimator.

Notice that in the case of **binary classification** ($y \in \{0, 1\}$), the probability density $p(y|\mathbf{x})$ in (2.5) reduces to

$$p(y|\mathbf{x}) = p(y = 0|\mathbf{x})\, \delta(y - 0) + p(y = 1|\mathbf{x})\, \delta(y - 1) \tag{2.6}$$

Where $\delta(x)$ is the Dirac delta distribution. The conditional criterion (2.5) can therefore be rewritten as

$$C[\widehat{y}|\mathbf{x}] = p(y = 0|\mathbf{x})\, \pounds[\widehat{y}; 0] + p(y = 1|\mathbf{x})\, \pounds[\widehat{y}; 1] \tag{2.7}$$

In the next section, we define a class of 'reasonable cost functions', and we derive the transformation that maps the output of the trained model $\widehat{y}^*$ to the a posteriori probability $p(\omega_1|\mathbf{x}) = p(y = 1|\mathbf{x}) = E_y\{y|\mathbf{x}\}$.

3. Mapping the output of the trained model to the a posteriori probability (binary output case)

3.1. A class of reasonable cost functions

For training our classifier, we must choose a cost function that measures the discrepancy between the model's output and the observed desired output. For this purpose, we will consider the class of cost functions $\pounds[\widehat{y}; y]$ of the type

$$\begin{cases} \pounds[\widehat{y}; y] & = 0 \text{ if and only if } \widehat{y} = y \\ \pounds[\widehat{y}; y] & > 0 \text{ if } \widehat{y} \neq y \\ \pounds[\widehat{y}; y] & \text{is twice continuously differentiable} \\ & \text{in terms of all its arguments} \end{cases} \tag{3.1}$$

We also make the natural requirement that when the predicted value $\widehat{y}$ moves away from the desired value y, the cost $\pounds[\widehat{y}; y]$ increases. Symmetrically, the cost $\pounds[\widehat{y}; y]$ should decrease when the predicted value $\widehat{y}$ approaches the desired value y. This implies that

$$\frac{\partial \pounds[\widehat{y}; y]}{\partial \widehat{y}} \text{ is } \begin{cases} > 0 & \text{if } \widehat{y} > y \\ < 0 & \text{if } \widehat{y} < y \end{cases} \tag{3.2}$$

and, together with (3.1), that

$$\left.\frac{\partial \pounds[\widehat{y}; y]}{\partial \widehat{y}}\right|_{\widehat{y}=y} = 0 \tag{3.3}$$

Finally, we also assume that $\pounds[\widehat{y}; y]$ depends on $\mathbf{x}$ only through the variable $\widehat{y}$.

Equations (3.1), (3.2) and (3.3) define the class of **reasonable cost functions** we will be working with. Some examples of such cost functions are provided in section 5.

3.2. Minimizing the criterion

Suppose now that we choose to train a sufficiently powerful model with one of these reasonable cost functions. This means that we pick up the model parameters $\mathbf{w}^*$ that minimize the performance criterion defined by equation (2.3), or equivalently the conditional criterion (2.7), in the binary classification case. The conditional criterion $C[\widehat{y}|\mathbf{x}]$ is therefore minimized for some optimal value $\widehat{y}^*(\mathbf{x}) = N[\mathbf{x}, \mathbf{w}^*]$ – since we assume that the model is "perfect", optimizing with respect to $\mathbf{w}$ is equivalent to optimizing with respect to $\widehat{y}$. This value, $\widehat{y}^*$, is the optimal output with respect to the criterion $C[\widehat{y}|\mathbf{x}]$ defined by (2.7). This means that the following standard optimality conditions must hold:

$$\left.\frac{\partial C[\widehat{y}|\mathbf{x}]}{\partial \widehat{y}}\right|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})} = 0 \tag{3.4}$$

$$\left.\frac{\partial^2 C[\widehat{y}|\mathbf{x}]}{\partial \widehat{y}^2}\right|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})} > 0 \tag{3.5}$$

for every $\mathbf{x}$. The model therefore supplies, after training, the output $\widehat{y}^*(\mathbf{x})$ representing some 'degree of plausibility' of the event $y = 1$.

We can easily show that the minimum of $C[\widehat{y}|\mathbf{x}]$ lies in the interval $[0, 1]$ ($\in [0, 1]$). Indeed, from (2.7),

$$\frac{\partial C[\widehat{y}|\mathbf{x}]}{\partial \widehat{y}} = p(y = 0|\mathbf{x}) \frac{\partial \pounds[\widehat{y}; 0]}{\partial \widehat{y}} + p(y = 1|\mathbf{x}) \frac{\partial \pounds[\widehat{y}; 1]}{\partial \widehat{y}}$$

and since $\partial \pounds[\widehat{y}|\mathbf{x}]/\partial \widehat{y} > 0$ when $\widehat{y} > y$ (3.2), for $\widehat{y} > 1$,

$$\begin{aligned} \frac{\partial C[\widehat{y}|\mathbf{x}]}{\partial \widehat{y}} &= p(y = 0|\mathbf{x}) \frac{\partial \pounds[\widehat{y}; 0]}{\partial \widehat{y}} + p(y = 1|\mathbf{x}) \frac{\partial \pounds[\widehat{y}; 1]}{\partial \widehat{y}} \\ &> 0 \text{ for } \widehat{y} > 1 \end{aligned}$$

so that $C[\widehat{y}|\mathbf{x}]$ is continuously increasing when $\widehat{y} > 1$ and $\widehat{y}$ increases above 1.

Symmetrically, we can show in a similar manner that $C[\widehat{y}|\mathbf{x}]$ is continuously increasing when $\widehat{y} < 0$ and $\widehat{y}$ decreases below 0 ($\partial C[\widehat{y}|\mathbf{x}]/\partial\widehat{y} < 0$ for $\widehat{y} < 0$).

The minimum of $C[\widehat{y}|\mathbf{x}]$ therefore $\in [0,1]$, and the fact that the output of the model $\in [0,1]$ ($0 \leq \widehat{y} \leq 1$; see (2.1)) is not a restriction at all, since the minimum is always attainable (it lies in $[0,1]$).

3.3. The mapping to a posteriori probabilities

Now that we have trained our model by optimizing the criterion $C[\widehat{y}|\mathbf{x}]$, the model provides as output $\widehat{y}^*(\mathbf{x})$ verifying (3.4).

In the appendix, we show that there always exists a transformation $f(\widehat{y}^*)$ that maps the model's optimal output $\widehat{y}^*(\mathbf{x})$ to the a posteriori probability $p(\omega_1|\mathbf{x}) = p(y = 1|\mathbf{x})$. This transformation is

$$\boxed{f(\widehat{y}^*) = \frac{\pounds'[\widehat{y}^*; 0]}{\pounds'[\widehat{y}^*; 0] - \pounds'[\widehat{y}^*; 1]}} \tag{3.6}$$

where $\pounds'[\widehat{y}^*; 0] = \partial\pounds[\widehat{y}; 0]/\partial\widehat{y}|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})}$ and $\pounds'[\widehat{y}^*; 1] = \partial\pounds[\widehat{y}; 1]/\partial\widehat{y}|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})}$. (3.6) can be rewritten in symmetric form:

$$f(\widehat{y}^*) = \frac{1}{1 - \dfrac{\pounds'[\widehat{y}^*; 1]}{\pounds'[\widehat{y}^*; 0]}} \tag{3.7}$$

This mapping transforms the optimal output of the model to the a posteriori probability

$$f(\widehat{y}^*(\mathbf{x})) = p(y = 1|\mathbf{x}) = p(\omega_1|\mathbf{x}) \tag{3.8}$$

Moreover, we also show in the appendix that if (3.5) holds for every $\widehat{y}^* \in [0,1]$, the mapping is one-to-one.

More precisely, in the Appendix A, we show that if the model has been trained by optimizing the criterion (it supplies optimal values $\widehat{y}^*$ verifying (3.4)) and if there exists a mapping that transforms the output of the model to the a posteriori probabilities (3.8), then this mapping is given by (3.6).

In the Appendix B, we show that if the model has been trained by optimizing the criterion (3.4) and we transform the model's output $\widehat{y}^*$ by (3.6), then the result of the mapping is $p(y = 1|\mathbf{x})$, the a posteriori probability of observing $y = 1$ conditional on $\mathbf{x}$ (equation 3.8).

Finally, in the Appendix C, we show that a second-order condition (3.5) holding for every $\widehat{y}^* \in [0,1]$ is equivalent to a strictly monotonic increasing $f(\widehat{y}^*)$ on $[0,1]$. In this case, the mapping is one-to-one, and the conditional criterion has only one global minimum (no local minimum). On the contrary, a non-monotonic

increasing mapping (i.e. the function $f(\widehat{y}^*)$ is stationary or decreasing on some interval $\in]0,1[$) is associated with multiple local minima of the conditional criterion, for some value of $p(y=1|\mathbf{x})$. We should therefore restrict the class of reasonable cost functions to those that have a strictly monotonic increasing mapping $f(\widehat{y}^*)$.

It is easy to verify that (3.6) is a function that maps the interval $[0,1]$ on $[0,1]$. Indeed, by examining (3.6), from (3.2) and the fact that $0 \leq \widehat{y}^* \leq 1$ (2.1), we easily find that $(\pounds'[\widehat{y}^*;0] - \pounds'[\widehat{y}^*;1]) > 0$ and $\pounds'[\widehat{y}^*;0] > 0$, so that $f(\widehat{y}^*) \geq 0$ for $\widehat{y}^* \in [0,1]$. Moreover, from (3.2), (3.3), and the fact that $f(\widehat{y}^*)$ is continuous, we deduce that $f(0)=0$, $f(1)=1$ and that $0 \leq f(\widehat{y}^*) \leq 1$ (the equation $f(\widehat{y}^*)=0$ has only one solution, $\widehat{y}^*=0$, on $[0,1]$; similarly, $f(\widehat{y}^*)=1$ has only one solution, $\widehat{y}^*=1$, so that $f(\widehat{y}^*)$ remains in $[0,1]$). The transformation $f(\widehat{y}^*)$ is therefore a function that maps the interval $[0,1]$ on $[0,1]$ (see section 5 for examples of mappings).

A remarkable property of (3.6) is the fact that the mapping only depends on the cost function $\pounds$ and, in particular, does not depend on $p(y|\mathbf{x})$. Moreover, we can easily show that if the cost function verifies the conditions that lead to the estimation of the a posteriori probability (stated in [3] and reproduced in [9]), the mapping reduces to $f(\widehat{y}^*) = \widehat{y}^*$.

A consequence of these results is that we are free to choose any reasonable cost function in order to train the classification model. If we need the a posteriori probabilities, we compute the mapping (3.6) in order to obtain an approximation of the Bayesian a posteriori probabilities.

Notice, however, that all our results are essentially asymptotic, and that issues regarding estimation from finite data sets are not addressed.

4. Some results for the multi-output case

All the previously derived results concern the binary output case. In this section, we will discuss the multi-output case, for which necessary conditions for obtaining a mapping to the a posteriori probabilities will be derived. However, the obtained results will be far less general than for the binary case.

In the multi-output case, we will consider that, for each training pattern $\mathbf{x}_k$, there is a corresponding **desired output vector** $\mathbf{y}_k$, where each $\mathbf{y}_k$ is associated to one of n mutually exclusive classes. That is, $\mathbf{y}_k$ will indicate the **class label** $\in \{\omega_1, \ldots, \omega_n\}$ of the observation $\mathbf{x}_k$. Each class label ω_i will be represented numerically by an **indicator vector** $\mathbf{e}_i$: if the observation $\mathbf{x}_k$ of the training set is assigned to the class label ω_i, then $\mathbf{y}_k = \mathbf{e}_i = [\underset{1}{0}, \ldots, \underset{i-1}{0}, \underset{i}{1}, \underset{i+1}{0}, \ldots, \underset{n}{0}]^{\mathrm{T}}$.

Correspondingly, the neural network provides a predicted value vector as output:

$$\widehat{\mathbf{y}}_k = N[\mathbf{x}_k, \mathbf{w}] \tag{4.1}$$

with $\widehat{\mathbf{y}}_k = [\widehat{y}_1(k), \widehat{y}_2(k), \ldots, \widehat{y}_n(k)]^{\mathrm{T}}$.We will assume that the outputs of the neural network sum to one ($\sum_{i=1}^{n} \widehat{y}_i = 1$), as it is often the case for classification models (see for example the case of a softmax nonlinearity [1], or a logistic regression model [2]). This means that the output vector $\widehat{\mathbf{y}}$ has only $n-1$ degrees of freedom, and can be represented by $\widehat{\mathbf{y}} = [\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_{n-1}, 1 - \sum_{i=1}^{n-1} \widehat{y}_i]^{\mathrm{T}}$.

Now, notice, as a particular case, that the mapping (3.6) can be applied to multi-output classifiers, provided that they are trained with a cost function which is a sum of individual scores, each score depending only on one output.

In full generality, for training the model, we will consider the class of cost functions $\pounds[\widehat{\mathbf{y}}; \mathbf{y}]$ of the type

$$\begin{cases} \pounds[\widehat{\mathbf{y}}; \mathbf{y}] & = 0 \text{ if and only if } \widehat{\mathbf{y}} = \mathbf{y} \\ \pounds[\widehat{\mathbf{y}}; \mathbf{y}] & > 0 \text{ if } \widehat{\mathbf{y}} \neq \mathbf{y} \\ \pounds[\widehat{\mathbf{y}}; \mathbf{y}] & \text{is twice continuously differentiable} \\ & \text{in terms of all its arguments} \end{cases} \tag{4.2}$$

By following the same steps as in section 2, the conditional criterion can be written as

$$C[\widehat{\mathbf{y}}|\mathbf{x}] = \int \pounds[\widehat{\mathbf{y}}; \mathbf{y}]\, p(\mathbf{y}|\mathbf{x})\, d\mathbf{y} = E_{\mathbf{y}}\{\pounds[\widehat{\mathbf{y}}; \mathbf{y}]|\mathbf{x}\} \tag{4.3}$$

In the classification case, the conditional criterion reduces to

$$C[\widehat{\mathbf{y}}|\mathbf{x}] = \sum_{j=1}^{n} p(\mathbf{y} = \mathbf{e}_j|\mathbf{x})\, \pounds[\widehat{\mathbf{y}}; \mathbf{e}_j] \tag{4.4}$$

A necessary set of equations for $\widehat{\mathbf{y}}^*$ to be an optimum of the criterion is given by

$$\left.\frac{\partial C[\widehat{\mathbf{y}}|\mathbf{x}]}{\partial \widehat{y}_i}\right|_{\widehat{\mathbf{y}}(\mathbf{x}) = \widehat{\mathbf{y}}^*(\mathbf{x})} = 0, \text{ for } i = 1 \ldots n-1 \tag{4.5}$$

Notice that there are only $n-1$ equations since we replaced $\widehat{y}_n$ by $\left(1 - \sum_{i=1}^{n-1} \widehat{y}_i\right)$.

In the Appendix D, we show that if there exists a mapping of the outputs of the model to the a posteriori probabilities

$$f_i(\widehat{\mathbf{y}}^*(\mathbf{x})) = p(\mathbf{y} = \mathbf{e}_i|\mathbf{x}) = p(\omega_i|\mathbf{x}), \text{ for } i = 1 \ldots n-1, \tag{4.6}$$

this mapping is provided by solving the following system of $n-1$ equations in terms of the $f_j(\widehat{\mathbf{y}}^*)$

$$\boxed{\sum_{j=1}^{n-1} \left[1 - \frac{\pounds_i'[\widehat{\mathbf{y}}^*; \mathbf{e}_j]}{\pounds_i'[\widehat{\mathbf{y}}^*; \mathbf{e}_n]}\right] f_j(\widehat{\mathbf{y}}^*) = 1, \text{ for } i = 1 \ldots n-1} \tag{4.7}$$

where $\pounds'_i[\widehat{\mathbf{y}}^*; \mathbf{e}_j] = \partial \pounds[\widehat{\mathbf{y}}; \mathbf{e}_j]/\partial \widehat{y}_i|_{\widehat{\mathbf{y}}(\mathbf{x})=\widehat{\mathbf{y}}^*(\mathbf{x})}$, and $f_n(\widehat{\mathbf{y}}^*) = 1 - \sum_{i=1}^{n-1} f_i(\widehat{\mathbf{y}}^*)$.

However, we were not able to provide sufficient conditions for the multi-output case. Indeed, several conditions should be checked before being able to state that these transformations exist and map the outputs to the a posteriori probabilities:

- After having minimized the criterion, we cannot be sure that the output values $\widehat{y}_i^* \in [0, 1]$;
- We should check that the system of equations (4.7) has indeed a solution;
- For $\widehat{\mathbf{y}}^*$ to be a minimum of $C[\widehat{\mathbf{y}}|\mathbf{x}]$, the matrix of second-order derivatives should be definite positive.

For the rather general cost function definition that we defined, these conditions are quite difficult to assess, and should be verified on a case-by-case basis, for the cost function being used.

5. Some examples

In this section, we provide examples of mappings to a posteriori probabilities. We consider six different cost functions, plot the corresponding mapping (3.6), and examine the effect of the mapping on the optimal output.

The six cost functions are:

$$\pounds[\widehat{y}; y] = \exp[y]\,(y - \widehat{y} - 1) + \exp[\widehat{y}] \tag{5.1}$$

$$\pounds[\widehat{y}; y] = (\widehat{y} - y)^4 \tag{5.2}$$

$$\pounds[\widehat{y}; y] = 1 - \exp\left[-(\widehat{y} - y)^2\right] \tag{5.3}$$

$$\pounds[\widehat{y}; y] = \log\left[1 + (\widehat{y} - y)^2\right] \tag{5.4}$$

$$\pounds[\widehat{\mathbf{y}}; \mathbf{y}] = \log\left[1 + \|\widehat{\mathbf{y}} - \mathbf{y}\|^2\right] \tag{5.5}$$

$$\pounds[\widehat{\mathbf{y}}; \mathbf{y}] = \exp\left[\|\widehat{\mathbf{y}} - \mathbf{y}\|^2\right] + \exp\left[-\|\widehat{\mathbf{y}} - \mathbf{y}\|^2\right] - 2 \tag{5.6}$$

These cost functions are displayed in table 5.1 and the corresponding mappings $f(\widehat{y}^*)$ provided by equation (3.6) and (4.7) are displayed in table 5.2. The first four cost functions ((5.1)-(5.4)) illustrate the binary output case; the two last cost functions ((5.5), (5.6)) illustrate a 3-output problem. In the later case (two last graphs of table 5.2), we show the mapping $f_1(\widehat{y}_1^*, \widehat{y}_2^*, \widehat{y}_3^*)$ with $\widehat{y}_1^* \in [0, 0.8]$, $\widehat{y}_2^* = 0.2$ and $\widehat{y}_3^* = (1 - \widehat{y}_2^* - \widehat{y}_3^*)$.

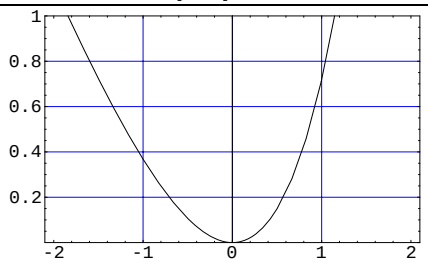
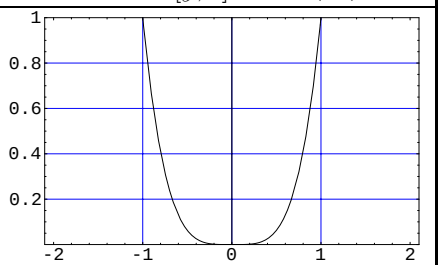
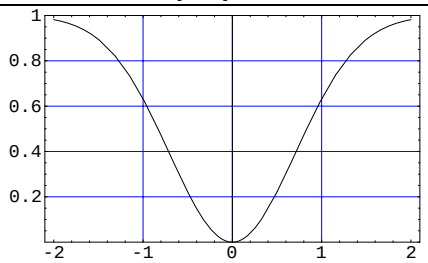
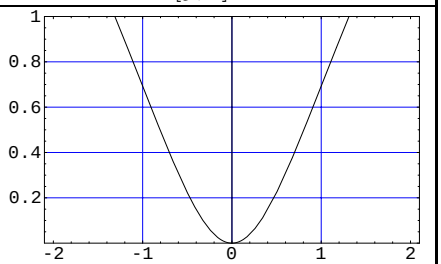
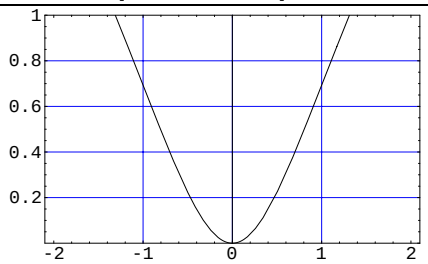
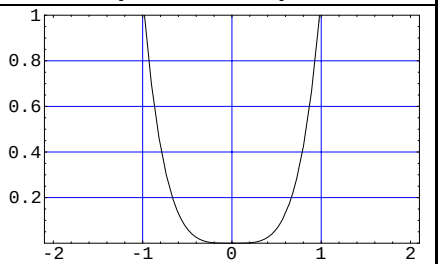
Graph of $\mathcal{L}[\widehat{y};0]$ for cost (5.1)	Graph of $\mathcal{L}[\widehat{y};0]$ for cost (5.2)
	
Graph of $\mathcal{L}[\widehat{y};0]$ for cost (5.3)	Graph of $\mathcal{L}[\widehat{y};0]$ for cost (5.4)
	
Graph of $\mathcal{L}[\widehat{y}_1,0,0;0,0,0]$ for cost (5.5)	Graph of $\mathcal{L}[\widehat{y}_1,0,0;0,0,0]$ for cost (5.6)
	

Table 5.1: Graph of the six cost functions. The first four functions illustrate the binary case; the two last ones illustrate the multi-output case (3 outputs).

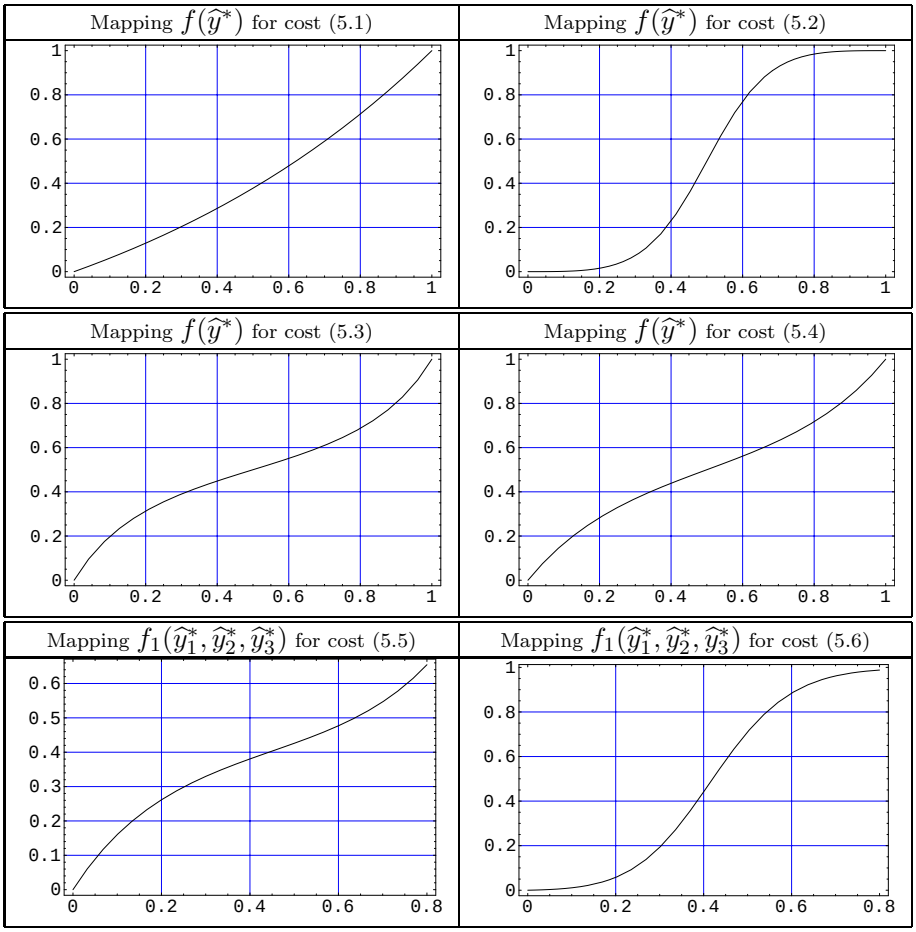


Table 5.2: Graph of the mapping to the a posteriori probabilities, for the six cost functions, as provided by equation (3.6) (binary case) and (4.7) (multi-output case).

By using the Mathematica software [11], for each of the first four cost functions ((5.1)-(5.4); binary output case), we compute the minimum $\widehat{y}^*$ of the criterion

$$\begin{aligned} C[\widehat{y}|\mathbf{x}] &= p(y=0|\mathbf{x})\,\pounds[\widehat{y};0] + p(y=1|\mathbf{x})\,\pounds[\widehat{y};1] \\ &= (1-p(y=1|\mathbf{x}))\,\pounds[\widehat{y};0] + p(y=1|\mathbf{x})\,\pounds[\widehat{y};1] \end{aligned} \tag{5.7}$$

for different values of $p(y=1|\mathbf{x})$ ranging from 0 to 1, illustrating all the potential situations tha can occur (table 5.3, plain line). These are the optimal outputs of the model corresponding to different class distributions $p(y=1|\mathbf{x})$ that can be encountered in a binary classification problem. Notice that $\mathbf{x}$ does not play any role here since all our probability densities are conditionned on $\mathbf{x}$.

Then, we transform the output $\widehat{y}^*$ by using the mapping $f(\widehat{y}^*)$ (3.6) and plot the results in terms of $p(y=1|\mathbf{x})$ (table 5.3, dash line). We clearly observe that the transformed output is mapped on the a posteriori probability ($f(\widehat{y}^*) = p(y=1|\mathbf{x})$).

For the multi-output case (two last cost functions (5.5), (5.6)), we plot the output $\widehat{y}_1^*$ before remapping (y-axis,plain line) and after remapping by $f_1(\widehat{y}_1^*, \widehat{y}_2^*, \widehat{y}_3^*)$ (see (4.7)) (y-axis, dash line), in function of the a posteriori probability $p(\mathbf{y}=\mathbf{e}_1|\mathbf{x})$ (x-axis), for values of $p(\mathbf{y}=\mathbf{e}_1|\mathbf{x}) \in [0, 0.8]$, $p(\mathbf{y}=\mathbf{e}_2|\mathbf{x}) = 0.2$, $p(\mathbf{y}=\mathbf{e}_3|\mathbf{x}) = 1 - p(\mathbf{y}=\mathbf{e}_1|\mathbf{x}) - p(\mathbf{y}=\mathbf{e}_2|\mathbf{x})$ (see table 5.3).

6. Conclusion

In this paper, we provide a straightforward proof of an important, but nevertheless little known, result that was published in 1982 by Lindley [4] in the framework of subjective probability theory. Lindley's result, when reformulated in the machine learning/pattern recognition context, puts new lights on the probabilistic interpretation of the outputs of a trained classifier.

Roughly speaking, it says that, when training a classification model by minimizing a cost function, it is always possible to map the output of the model to the Bayesian a posteriori probabilities of the classes.

However, we must keep in mind that the results obtained in this paper are only valid if

- A minimum of the criterion is indeed reached after training, and
- The neural network is a 'sufficiently powerful model' that is able to approximate the optimal estimator to any degree of accuracy (perfect model matching).

Notice also that the results presented here are essentially asymptotic, and that issues regarding estimation from finite data sets are not addressed.

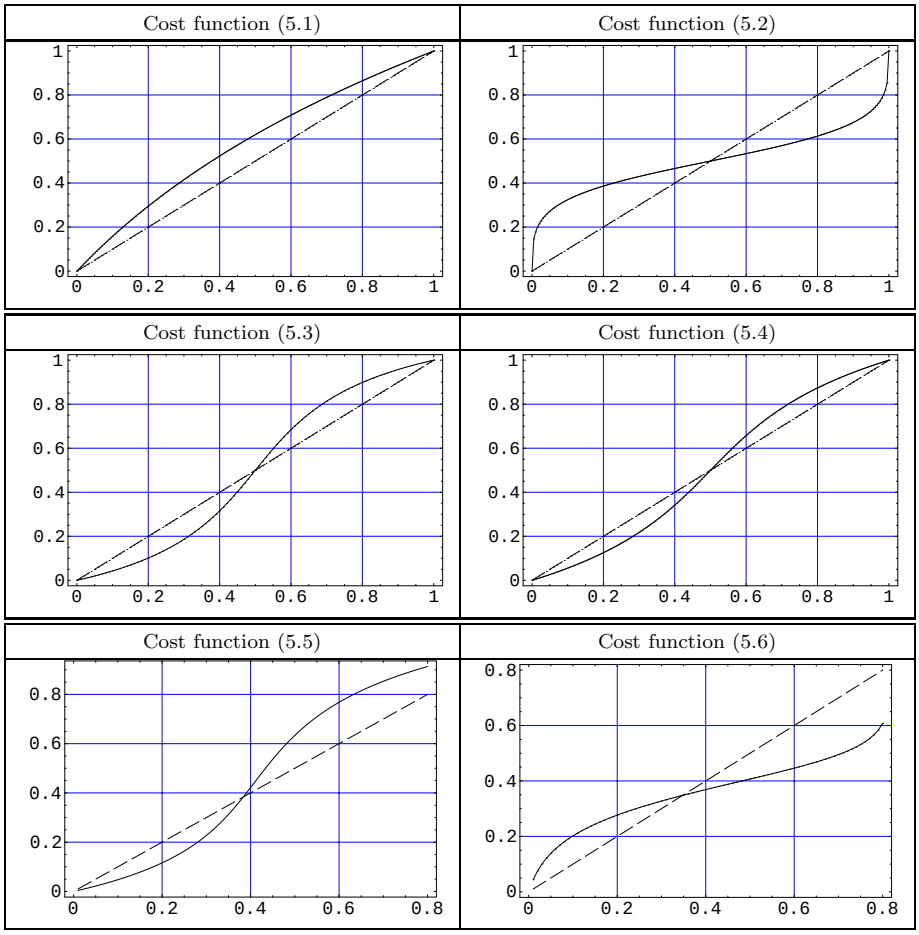


Table 5.3: Graphs of the optimal output obtained after minimization of the criterion before remapping (y-axis, plain line) and after remapping the output (y-axis, dash line), plotted in relation with different a posteriori probabilities (x-axis), for the six different cost functions. The applied mappings are shown in the table (5.3). The first four graphs are for binary models; the last two are for 3-output models. We clearly observe that the transformation maps the output of the model on the a posteriori probability of the class.

Acknowledgments

This work was partially supported by the project RBC-BR 216/4041 from the 'Région de Bruxelles-Capitale', and funding from the SmalS-MvM. Patrice Latinne is supported by a grant under an ARC (Action de Recherche Concertée) program of the Communauté Française de Belgique. We also thank the two anonymous reviewers for their pertinent and constructive remarks.

References

[1] Bishop C. (1995). "Neural networks for pattern recognition". Oxford University Press.

[2] Fomby T., Carter Hill R. & Johnson S. (1984). "Advanced econometric methods". Springer-Verlag.

[3] Hampshire J.B. & Pearlmutter B. (1990). "Equivalence proofs for multi-layer perceptron classifiers and the Bayesian discriminant function". In *Proceedings of the 1990 Connectionnist Models Summer School*, Touretzky D., Elman J., Sejnowski T. & Hinton G. (editors), Morgan Kaufmann, pp. 159-172.

[4] Lindley D. (1982). "Scoring rules and the inevitability of probability (with discussions)". *International Statistical Review*, 50, pp. 1-26.

[5] Richard M.D. & Lippmann R.P. (1991). "Neural network classifiers estimate Bayesian a posteriori probabilities". *Neural Computation*, 3, pp. 461-483.

[6] McCullagh P. & Nelder J.A. (1990) "Generalized linear models, 2nd ed". Chapman and Hall.

[7] Miller J.W., Goodman R. & Smyth P. (1991). "Objective functions for probability estimation". Proceedings of the *IEEE International Joint Conference on Neural Networks*, San Diego, pp. I-881-886.

[8] Miller J.W., Goodman R. & Smyth P. (1993). "On loss functions which minimize to conditional expected values and posterior probabilities". *IEEE Transactions on Information Theory*, IT-39 (4), pp. 1404-1408.

[9] Saerens M. (1996). "Non mean square error criteria for the training of learning machines". Proceedings of the 13th *International Conference on Machine Learning* (ICML), July 1996, Bari (Italy), pp. 427-434.

[10] Saerens M. (2000). "Building cost functions minimizing to some summary statistics". *IEEE Transactions on Neural Networks*, NN-11 (6), pp. 1263-1271.

[11] Wolfram S. (1999). "*The Mathematica Book, 4th ed.*". Wolfram Media & Cambridge University Press.

A. Appendix: If the model is trained by optimizing $C[\widehat{y}|\mathbf{x}]$ (equation 3.4), and if there exists a mapping that transforms the output of the model $\widehat{y}^*$ to the a posteriori probabilities (equation 3.8), then this mapping is provided by (3.6)

Let us recall the different hypothesis. After training, the criterion attains its optimal value at $\widehat{y}^*(\mathbf{x})$. Thus, from (2.7) and (3.4), we obtain

$$\left.\frac{\partial C[\widehat{y}|\mathbf{x}]}{\partial \widehat{y}}\right|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})} = \left.\frac{\partial \pounds[\widehat{y};0]}{\partial \widehat{y}}\right|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})} p(y=0|\mathbf{x}) + \left.\frac{\partial \pounds[\widehat{y};1]}{\partial \widehat{y}}\right|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})} p(y=1|\mathbf{x}) = 0 \tag{A.1}$$

Moreover, let us suppose that there exists a mapping that transforms the optimal output $\widehat{y}^*(\mathbf{x})$ to the a posteriori probabilities:

$$f(\widehat{y}^*) = p(y=1|\mathbf{x}) = p(\omega_1|\mathbf{x}) \tag{A.2}$$

with

$$p(y=0|\mathbf{x}) + p(y=1|\mathbf{x}) = 1 \tag{A.3}$$

By developing (A.1) and using (A.2)–(A.3), we easily obtain

$$\left.\frac{\partial \pounds[\widehat{y};0]}{\partial \widehat{y}}\right|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})} (1-f(\widehat{y}^*)) + \left.\frac{\partial \pounds[\widehat{y};1]}{\partial \widehat{y}}\right|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})} f(\widehat{y}^*) = 0 \tag{A.4}$$

from which we compute $f(\widehat{y}^*)$

$$f(\widehat{y}^*) = \frac{\pounds'[\widehat{y}^*;0]}{\pounds'[\widehat{y}^*;0] - \pounds'[\widehat{y}^*;1]} \tag{A.5}$$

where $\pounds'[\widehat{y}^*;0] = \partial \pounds[\widehat{y};0]/\partial \widehat{y}|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})}$ and $\pounds'[\widehat{y}^*;1] = \partial \pounds[\widehat{y};1]/\partial \widehat{y}|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})}$.

This shows that if there exists a mapping to the a posteriori probabilities, this mapping is provided by (3.6). ■

B. Appendix: If the model is trained by optimizing $C[\widehat{y}|\mathbf{x}]$ (equation 3.4), and we transform the model's output $\widehat{y}^*$ by (3.6), then the result of the mapping is the a posteriori probability defined by (3.8)

As in appendix A, let us consider a trained model (equation (3.4) is verified). From (2.7),

$$\left.\frac{\partial C[\widehat{y}|\mathbf{x}]}{\partial \widehat{y}}\right|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})} = \pounds'[\widehat{y}^*;0]\,p(y=0|\mathbf{x}) + \pounds'[\widehat{y}^*;1]\,p(y=1|\mathbf{x}) = 0 \qquad \text{(B.1)}$$

where $\pounds'[\widehat{y}^*;0] = \partial\pounds[\widehat{y};0]/\partial\widehat{y}|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})}$ and $\pounds'[\widehat{y}^*;1] = \partial\pounds[\widehat{y};1]/\partial\widehat{y}|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})}$. From (B.1), we obtain

$$\frac{\pounds'[\widehat{y}^*;0]}{\pounds'[\widehat{y}^*;0] - \pounds'[\widehat{y}^*;1]} = p(y=1|\mathbf{x}) \qquad \text{(B.2)}$$

If we apply the mapping

$$f(\widehat{y}^*) = \frac{\pounds'[\widehat{y}^*;0]}{\pounds'[\widehat{y}^*;0] - \pounds'[\widehat{y}^*;1]} \qquad \text{(B.3)}$$

we find

$$f(\widehat{y}^*) = p(y=1|\mathbf{x}) \qquad \text{(B.4)}$$

Since we require that the cost function is twice differentiable (3.1), the mapping (B.3) always exists; it transforms the optimal output $\widehat{y}^*$ to the a posteriori probability $p(y=1|\mathbf{x})$. ■

C. Appendix: A conditional criterion $C[\widehat{y}|\mathbf{x}]$ (2.4) having only one global minimum (no local minimum) for every possible $p(y=1|\mathbf{x})$ is equivalent to a strictly monotonic increasing mapping $f(\widehat{y}^*)$ (3.6)

Notice that the requirements on the cost function (3.1)–(3.3) do not guarantee that the criterion has only one global minimum (no local minimum). Let us consider that $C[\widehat{y}|\mathbf{x}]$ is already optimized, and therefore (3.4) is verified. From appendix A and B, this means that the optimum of $C[\widehat{y}|\mathbf{x}]$, denoted by $\widehat{y}^*$, is such that

$$p(y=1|\mathbf{x}) = \frac{\pounds'[\widehat{y}^*;0]}{\pounds'[\widehat{y}^*;0] - \pounds'[\widehat{y}^*;1]} \qquad \text{(C.1)}$$

In this appendix, we are interested in the second-order properties of the criterion. For $\widehat{y}^*$ to be a minimum, the second-order condition (3.5) should be verified

in addition to (3.4). Let us compute the second-order derivative of $C[\widehat{y}|\mathbf{x}]$. From (2.7), we have

$$\left.\frac{\partial^2 C[\widehat{y}|\mathbf{x}]}{\partial \widehat{y}^2}\right|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})} = (1-p(y=1|\mathbf{x}))\ \pounds''[\widehat{y}^*;0] + p(y=1|\mathbf{x})\ \pounds''[\widehat{y}^*;1] \quad \text{(C.2)}$$

Where $\pounds''[\widehat{y}^*;0] = \left.\partial^2 \pounds[\widehat{y};0]/\partial \widehat{y}^2\right|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})}$ and $\pounds''[\widehat{y}^*;1] = \left.\partial^2 \pounds[\widehat{y};1]/\partial \widehat{y}^2\right|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})}$. Since we are at an optimum of $C[\widehat{y}|\mathbf{x}]$, we can substitute $p(y=1|\mathbf{x})$ by (C.1) in (C.2). We obtain

$$\left.\frac{\partial^2 C[\widehat{y}|\mathbf{x}]}{\partial \widehat{y}^2}\right|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})} = \frac{\pounds'[\widehat{y}^*;0]\ \pounds''[\widehat{y}^*;1] - \pounds'[\widehat{y}^*;1]\ \pounds''[\widehat{y}^*;0]}{(\pounds'[\widehat{y}^*;0] - \pounds'[\widehat{y}^*;1])} \quad \text{(C.3)}$$

Now, let us also compute the first derivative of the mapping $f(\widehat{y}^*)$ (equation (3.6))

$$\frac{\partial f(\widehat{y}^*)}{\partial \widehat{y}^*} = \frac{\pounds'[\widehat{y}^*;0]\ \pounds''[\widehat{y}^*;1] - \pounds'[\widehat{y}^*;1]\ \pounds''[\widehat{y}^*;0]}{(\pounds'[\widehat{y}^*;0] - \pounds'[\widehat{y}^*;1])^2} \quad \text{(C.4)}$$

Since $\widehat{y}^* \in [0,1]$, from (3.2), $(\pounds'[\widehat{y}^*;0] - \pounds'[\widehat{y}^*;1]) > 0$. Therefore, by comparing (C.3) and (C.4), we observe that $\left.\partial^2 C[\widehat{y}|\mathbf{x}]/\partial \widehat{y}^2\right|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})} > 0$ is equivalent to $\partial f(\widehat{y}^*)/\partial \widehat{y}^* > 0$ (the numerators are the same and the denominators are positive). This means that a monotonic increasing $f(\widehat{y}^*)$ $(\partial f(\widehat{y}^*)/\partial \widehat{y}^* > 0)$ for every $\widehat{y}^* \in [0,1]$ is equivalent to the fact that the conditional criterion is a minimum at every $\widehat{y}^* \in [0,1]$ $(\left.\partial^2 C[\widehat{y}|\mathbf{x}]/\partial \widehat{y}^2\right|_{\widehat{y}(\mathbf{x})=\widehat{y}^*(\mathbf{x})} > 0)$.

Let us now examine what happens when the conditional criterion has a local minimum. Suppose that $C[\widehat{y}|\mathbf{x}]$ has two minima at $\widehat{y}^*$ and $\widehat{y}'^*$ (with $\widehat{y}^* < \widehat{y}'^*$), for the same $p(y=1|\mathbf{x})$. In this case, since $C[\widehat{y}|\mathbf{x}]$ is differentiable, it must pass through a maximum $\widehat{y}_{\max}$ located between the two minima $\widehat{y}^* < \widehat{y}_{\max} < \widehat{y}'^*$. For this maximum, we have $\left.\partial^2 C[\widehat{y}|\mathbf{x}]/\partial \widehat{y}^2\right|_{\widehat{y}=\widehat{y}_{\max}} < 0$ which is equivalent to $\left.\partial f(\widehat{y})/\partial \widehat{y}\right|_{\widehat{y}=\widehat{y}_{\max}} < 0$, and therefore a decreasing $f(\widehat{y}^*)$ on some interval including $\widehat{y}_{\max}$. This indicates that a decreasing $f(\widehat{y}^*)$ on some interval is associated to local minima of the conditional criterion.

This shows that conditional criterion (2.4) having only one global minimum (no local minimum) for every possible $p(y=1|\mathbf{x})$ is equivalent to a strictly monotonic increasing mapping (3.6). ■

D. Appendix: multi-output case. If the model is trained by optimizing $C[\widehat{\mathbf{y}}|\mathbf{x}]$ (equation 4.5), and if there exists a mapping that transforms the output of the model $\widehat{\mathbf{y}}^*$ to the a posteriori probabilities (equation 4.6), then this mapping is obtained by solving a system of $n-1$ linear equations (4.7)

After training, the criterion attains its optimal value at $\widehat{\mathbf{y}}^*(\mathbf{x})$. Thus, from (4.4) and (4.5), we obtain

$$\begin{aligned} \left.\frac{\partial C[\widehat{\mathbf{y}}|\mathbf{x}]}{\partial \widehat{y}_i}\right|_{\widehat{\mathbf{y}}(\mathbf{x})=\widehat{\mathbf{y}}^*(\mathbf{x})} &= \sum_{j=1}^{n} p(\mathbf{y}=\mathbf{e}_j|\mathbf{x}) \left.\frac{\partial \mathcal{L}[\widehat{\mathbf{y}};\mathbf{e}_j]}{\partial \widehat{y}_i}\right|_{\widehat{\mathbf{y}}(\mathbf{x})=\widehat{\mathbf{y}}^*(\mathbf{x})} \\ &= 0, \text{ for } i=1\ldots n-1 \end{aligned} \tag{D.1}$$

Moreover, let us suppose that there exists a mapping that transforms the optimal output vector $\widehat{\mathbf{y}}^*(\mathbf{x})$ to the a posteriori probabilities:

$$f_i(\widehat{\mathbf{y}}^*(\mathbf{x})) = p(\mathbf{y}=\mathbf{e}_i|\mathbf{x}) = p(\omega_i|\mathbf{x}) \tag{D.2}$$

with

$$\sum_{i=1}^{n} f_i(\widehat{\mathbf{y}}^*(\mathbf{x})) = \sum_{i=1}^{n} p(\mathbf{y}=\mathbf{e}_i|\mathbf{x}) = 1 \tag{D.3}$$

By using (D.2) and (D.1), we easily obtain

$$\sum_{j=1}^{n} f_j(\widehat{\mathbf{y}}^*) \left.\frac{\partial \mathcal{L}[\widehat{\mathbf{y}};\mathbf{e}_j]}{\partial \widehat{y}_i}\right|_{\widehat{\mathbf{y}}(\mathbf{x})=\widehat{\mathbf{y}}^*(\mathbf{x})} = 0, \text{ for } i=1\ldots n-1 \tag{D.4}$$

Let us define $\mathcal{L}'_i[\widehat{\mathbf{y}}^*;\mathbf{e}_j] = \partial \mathcal{L}[\widehat{\mathbf{y}};\mathbf{e}_j]/\partial \widehat{y}_i|_{\widehat{\mathbf{y}}(\mathbf{x})=\widehat{\mathbf{y}}^*(\mathbf{x})}$. By further using (D.3), we rewrite (D.4) as

$$\sum_{j=1}^{n-1} f_j(\widehat{\mathbf{y}}^*)\, \mathcal{L}'_i[\widehat{\mathbf{y}}^*;\mathbf{e}_j] + \left[1 - \sum_{j=1}^{n-1} f_j(\widehat{\mathbf{y}}^*)\right] \mathcal{L}'_i[\widehat{\mathbf{y}}^*;\mathbf{e}_n] = 0, \text{ for } i=1\ldots n-1 \tag{D.5}$$

By rearranging the terms, we obtain

$$\sum_{j=1}^{n-1} \left[\mathcal{L}'_i[\widehat{\mathbf{y}}^*;\mathbf{e}_n] - \mathcal{L}'_i[\widehat{\mathbf{y}}^*;\mathbf{e}_j]\right] f_j(\widehat{\mathbf{y}}^*) = \mathcal{L}'_i[\widehat{\mathbf{y}}^*;\mathbf{e}_n], \text{ for } i=1\ldots n-1 \tag{D.6}$$

Or equivalently

$$\sum_{j=1}^{n-1} \left[1 - \frac{\mathcal{L}'_i[\widehat{\mathbf{y}}^*;\mathbf{e}_j]}{\mathcal{L}'_i[\widehat{\mathbf{y}}^*;\mathbf{e}_n]}\right] f_j(\widehat{\mathbf{y}}^*) = 1, \text{ for } i=1\ldots n-1 \tag{D.7}$$

This shows that if there exists a mapping to the a posteriori probabilities, this mapping is provided by solving (4.7).

However, for such general cost function definitions, it is difficult to assess if this solution exists and if it is indeed a minimum. ■
