LETTER **Communicated by Jose C. Principe**

Error Entropy in Classification Problems: A Univariate Data Analysis

Luís M. Silva
lmsilva@fe.up.pt
Carlos S. Felgueiras
casf@fe.up.pt
Instituto de Engenharia Biomédica, Laboratório Sinal e Imagem Biomédica, 4200-465, Porto, Portugal

Luís A. Alexandre
lfbaa@di.ubi.pt
Departamento de Informática, Universidade da Beira Interior, Covilhã, Portugal, and Instituto de Telecomunicações, Networks and Multimedia Group, Covilhã, Portugal

J. Marques de Sá
jmsa@fe.up.pt
Instituto de Engenharia Biomédica, Laboratório Sinal e Imagem Biomédica, 4200-465, Porto, Portugal, and Faculdade de Engenharia da Universidade do Porto, Departamento de Engenharia Electrotécnica e Computadores, 4200-465, Porto, Portugal

Entropy-based cost functions are enjoying a growing attractiveness in unsupervised and supervised classification tasks. Better performances in terms both of error rate and speed of convergence have been reported. In this letter, we study the principle of error entropy minimization (EEM) from a theoretical point of view. We use Shannon's entropy and study univariate data splitting in two-class problems. In this setting, the error variable is a discrete random variable, leading to a not too complicated mathematical analysis of the error entropy. We start by showing that for uniformly distributed data, there is equivalence between the EEM split and the optimal classifier. In a more general setting, we prove the necessary conditions for this equivalence and show the existence of class configurations where the optimal classifier corresponds to maximum error entropy. The presented theoretical results provide practical guidelines that are illustrated with a set of experiments with both real and simulated data sets, where the effectiveness of EEM is compared with the usual mean square error minimization.

1 Introduction

Entropy and related concepts of mutual information and Kulback-Leibler divergence have been used in learning systems (supervised or unsupervised) in several ways. The principle of minimum cross-entropy enunciated by Kullback (1959) was introduced as a powerful tool to build complete probability distributions when only partial knowledge is available.

The maximization of mutual information between input and output of a neural network (the Infomax principle) was introduced by Linsker (1988) as an unsupervised method that can be applied, for example, to feature extraction. Recently Principe, Xu, and Fisher (2000) proposed new approaches to the application of entropic criteria to learning systems. In particular, they proposed the minimization of Rényi's quadratic entropy of the error for regression, time series prediction, and feature extraction tasks (Erdogmus & Principe, 2000, 2002). The principle is as follows. Having an adaptive system (e.g., a neural network) with output variable Y and a target variable T, the error variable is measured as the difference between the target and the output of the system, $E = T - Y$. The minimization of error entropy implies a reduction on the expected information contained in the error, which leads to the maximization of the mutual information between the desired target and the system output (Erdogmus & Principe, 2000). This means that the classifier is learning the target variable.

Entropy-based cost functions, as functions of the probability density functions, reflect the global behavior of the error distribution; therefore, learning systems with entropic cost functions are expected to outperform those that use the popular mean square error (MSE) rule, which reflects only the second-order statistics of the error.

In this letter, we are concerned with the criterion of error entropy minimization (EEM) between the output of a classifier and the desired target. Santos, Alexandre, and Marques de Sá (2004) and Santos, Marques de Sá, Alexandre, and Sereno (2004) applied the EEM rule using Renyi's quadratic entropy to classification tasks, obtaining better results than with MSE rule. Silva, Marques de Sá, and Alexandre (2005) have also proposed the use of Shannon's entropy with the EEM principle; the results were also better than those obtained with MSE.

Despite the evidence provided by these experimental results, which suggests that EEM is an interesting alternative to the MSE principle, very little is known about the theoretical properties of EEM when applied to data classification, in terms of convergence to the optimal classifier, as well as whether Bayes error is attainable. This letter is meant as a contribution to the theoretical study of the EEM principle, using Shannon's entropy, in classification tasks. We analyze the case of univariate data splitting in two-class problems. We will use Shannon's formula (Shannon, 1948) for the entropy

of a discrete random variable X, H_X, taking N values with probability p_i

$$H_X = -\sum_{i=1}^{N} p_i \log p_i. \tag{1.1}$$

Despite the simplicity of the univariate data splitting model, this analysis will provide interesting insights and practical guidelines for the error entropy minimization rule.

The organization of the letter is as follows: In section 2 we introduce the univariate data splitting problem: In section 3 we analyze univariate EEM splits in the case of uniformly distributed data and show their convergence to the optimal classifier: In section 4 we present a more generalized analysis of univariate EEM splits and show the existence of situations where the optimal classifier corresponds to maximum error entropy. In section 5 we illustrate with simulated and real data the presented theoretical results: Finally, in section 6, we draw some conclusions and discuss future work.

2 The Univariate Split Problem

Let us consider the two-class classification problem with class-conditional distributions given by $F_t(x) = P(X \in]-\infty, x] | T = t)$, $t \in \{-1, 1\}$, where X and T are univariate input and target random variables, respectively, and $f_t(x)$ the corresponding probability density functions (pdf). The simplest possible linear discrimination rule corresponds to a classifier output, y, as

$$y = g(x) = \begin{cases} y', & x \leq x' \\ -y', & x > x' \end{cases}, \tag{2.1}$$

where x' is a data-splitting threshold and $y' \in \{-1, 1\}$ is a class label. The theoretic optimal rule corresponds to a split point x^* and class label y^* such that

$$(x^*, y^*) = \arg\min P(g(X) \neq T) \tag{2.2}$$

with minimum probability of error, P^*, given by

$$\begin{aligned} P^* = \inf\{ & I_{y'=-1}(pF_1(x') + q(1 - F_{-1}(x'))) \\ & + I_{y'=1}(p(1 - F_1(x')) + qF_{-1}(x'))\}, \end{aligned} \tag{2.3}$$

where $p = P(T = 1)$ and $q = P(T = -1)$ (the class priors). In equation 2.3, the first term inside braces corresponds to the situation where P^* is reached when $y' = -1$ is at the left of x'; the second term corresponds to swapping

the class labels. A split given by (x^*, y^*) is called a theoretical Stoller split (for details see Devroye, Gyorfi, & Lugosi, 1996).

We define the error variable $E = T - Y$, as the difference between the target and the classifier's output and notice that $E \in \{-2, 0, 2\}$[1].

What does it mean to minimize the error entropy in this situation? Does it also lead to the optimal solution for the class of linear threshold decision rules represented by equation 2.1?

As we are dealing with a discrete random variable, entropy is a concave function of the p_i in equation 1.1 (Kapur, 1993), where each p_i corresponds to the probability of E taking one of the values $\{-2, 0, 2\}$. These are precisely the probabilities of error P_t for each class $t \in \{-1, 1\}$ and the probability of correct classification $1 - \sum_t P_t$. Denoting $F_t(x')$ simply as F_t and considering from now on, without loss of generality, that $y' = -1$, one has

$$\begin{aligned} P_{-1} = P(E = -2) &= q\ (1 - F_{-1}) \\ P_1 = P(E = 2) &= p\ F_1 \\ 1 - P_{-1} - P_1 = P(E = 0) &= q\,F_{-1} + p\ (1 - F_1). \end{aligned} \tag{2.4}$$

Thus, the discrete entropy is

$$H_E = -[P_{-1} \log P_{-1} + P_1 \log P_1 + (1 - P_{-1} - P_1) \log (1 - P_{-1} - P_1)]. \tag{2.5}$$

In the following sections, we study the behavior of equation 2.5 as we vary x'. Section 3 is devoted to the case of uniform distributions, and the following sections consider the situation where the data distributions can be described in terms of continuous class-conditional density functions, where the following applies:

Theorem 1. *For continuous class-conditional density functions f_{-1} and f_1, the Stoller split occurs at an intersection of either $q f_{-1}$ with $p f_1$ or at $+\infty$ or $-\infty$.*

Proof. See section B.1.

3 EEM Splits for Uniform Distributions

Let us consider that the two classes have univariate uniform distributions,

$$f_{-1}(x) = \frac{1}{b-a} I_{[a,b]}(x) \qquad f_1(x) = \frac{1}{d-c} I_{[c,d]}(x), \tag{3.1}$$

[1] $E = -2$ and $E = 2$ means a misclassification for class $t = -1$ and $t = 1$, respectively. $E = 0$ means correct classification.

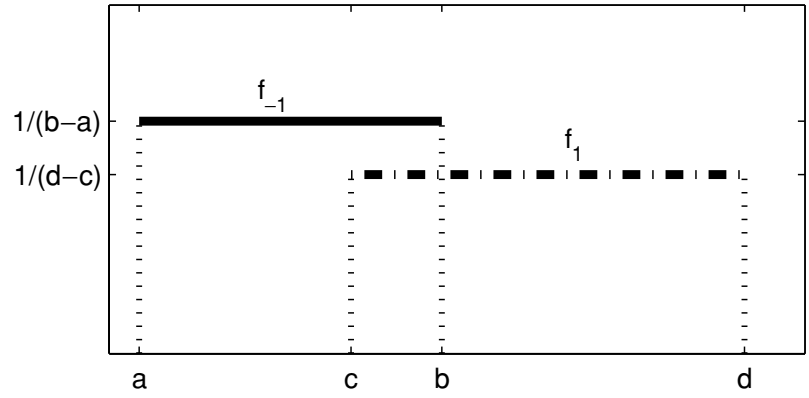

Figure 1: Schematic drawing of the simple problem of setting x' to classify two uniform overlapped classes.

where $I(x)$ is the indicator function. We first assume that the classes overlap, such that $a < c \leq b < d$. Figure 1 depicts this situation in terms of the density functions $f_t(x)$.

For this problem and making use of the formulas in equation 2.4, it is straightforward to compute H_E as in equation 2.5 for x' varying on the real line. Indeed, one has

$$H_E(x') = -\begin{cases} q\log q + 0\log 0 + p\log p, & x' < a \\ q\frac{b-x'}{b-a}\log\left(q\frac{b-x'}{b-a}\right) + 0\log 0 + \left(q\frac{x'-a}{b-a} + p\right)\log\left(q\frac{x'-a}{b-a} + p\right), & x' \in [a, c[\\ q\frac{b-x'}{b-a}\log\left(q\frac{b-x'}{b-a}\right) + p\frac{x'-c}{d-c}\log\left(p\frac{x'-c}{d-c}\right) & \\ \qquad + \left(q\frac{x'-a}{b-a} + p\frac{d-x'}{d-c}\right)\log\left(q\frac{x'-a}{b-a} + p\frac{d-x'}{d-c}\right), & x' \in [c, b[\\ 0\log 0 + p\frac{x'-c}{d-c}\log\left(p\frac{x'-c}{d-c}\right) + \left(q + p\frac{d-x'}{d-c}\right)\log\left(q + p\frac{d-x'}{d-c}\right), & x' \in [b, d[\\ 0\log 0 + p\log p + q\log q, & x' \geq d. \end{cases} \tag{3.2}$$

Figure 2 (dashed line) shows some examples for $p = 1/2$, $[a, b] = [0, 1]$ and different values of c and d.

First, one can see that although within each interval of x' (corresponding to the different cases above), H_E is a concave function, as a whole H_E is not concave. Second, whenever the overlap is nondegenerate (all figures of Figure 2 except 2c), we have two local minima located at the extremes of the

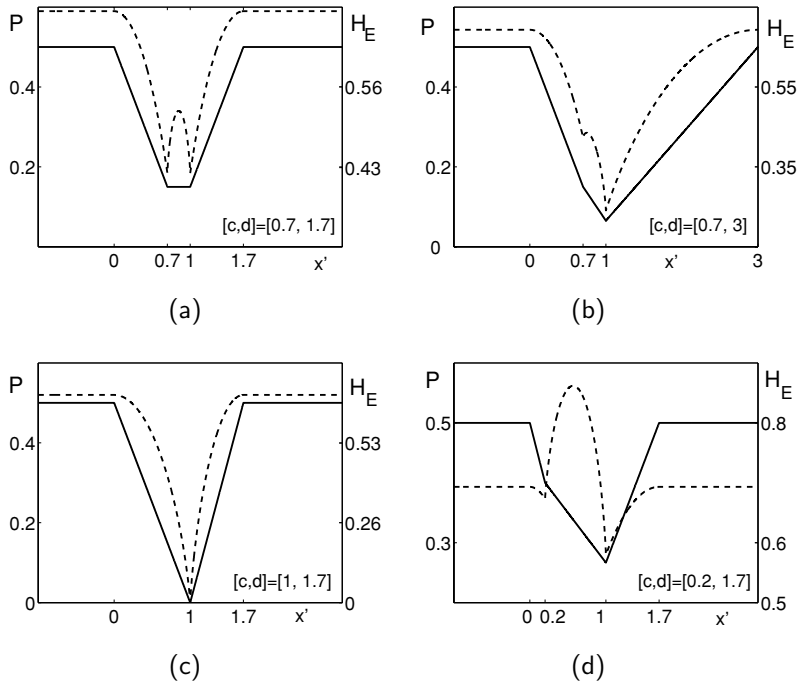

Figure 2: Shannon entropy (dashed line) and probability of error (solid line) plotted as functions of x'.

overlapped regions. A local maximum (global in some cases, as in Figure 2d), say, at x_0, is located within the overlapped region. If we have equal support for the two distributions (and equal priors), entropy is perfectly symmetric at x_0, and this is exactly the midpoint of the overlapped region (see Figure 2a). In the other cases, we have a local and a global minimum, and x_0 is deviated toward the former. Let us now determine the probability of error P for this example. Making use of the above expressions, we have

$$P(x') = P_{-1}(x') + P_1(x') = \begin{cases} q, & x' < a \\ q\,\dfrac{b-x'}{b-a}, & x' \in [a, c[\\ q\,\dfrac{b-x'}{b-a} + p\dfrac{x'-c}{d-c}, & x' \in [c, b[\\ p\,\dfrac{x'-c}{d-c}, & x' \in [b, d[\\ p, & x' \geq d. \end{cases} \tag{3.3}$$

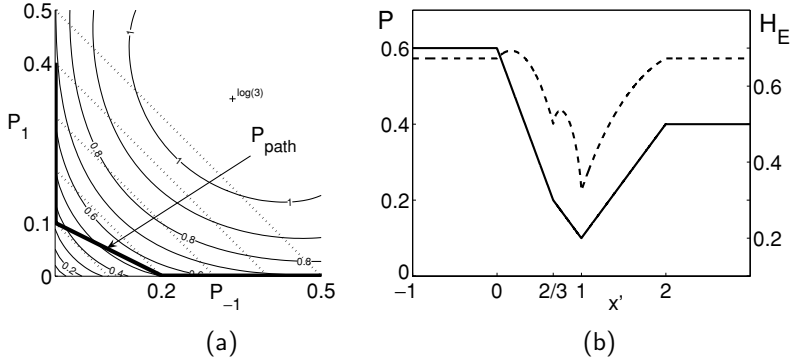

Figure 3: (a) Contour lines of H_E with a general path, P_{path}, produced by P. (b) P and H_E plotted as functions of x' for the path P_{path} in (a).

Figure 2 (solid line) plots P as a function of x' for the same values of a, b, c, and d. One can see that the global minimum of the error entropy corresponds to the theoretical Stoller split. In fact, for this problem, it also corresponds to the optimal decision in the Bayes sense. If we take the special case where $b - a = d - c$ (see Figure 2a), using the minimum probability of error criteria, we may locate x^* anywhere in $[c, b]$; for entropy, it is preferable to choose either $x^* = c$ or $x^* = b$. The reason is that the choice $x^* \in]c, b[$ increases the uncertainty or unstability of the system. At c or b, E takes only two values of $\{-2, 0, 2\}$; otherwise, E can assume every value in that set, which implies an increase in entropy. In other words, entropy prefers to classify correctly one class and leave all the errors to the other one.

Figure 2 can be easily reproduced for unequal priors, where the general behavior is the same. In fact, we can show that:

Theorem 2. *Suppose we have two overlapped uniform distributions as in equation 3.1 such that $a < c \leq b < d$. H_E and P have the same global minimum.*

Proof. Consider the $P_{-1} \times P_1$ plane. First, notice that a probability path, P_{path}, produced by P as in equation 3.3 is always composed by three linear segments: two along the axes connected by the remaining one (in some situations degenerated in one point). Second, notice that H_E, as a function of the probabilities, is concave and symmetric about the vertical plane $P_{-1} = P_1$. Therefore, the global minimum of H_E always coincides with the global minimum of the probability of error. The demonstration is illustrated in Figure 3a, where the contour lines of H_E are plotted as functions of P_{-1} and P_1. The solid line represents P_{path} (see Figure 3b plots P and H_E as functions of x' for this path), and the dashed lines are contours of equal probability.

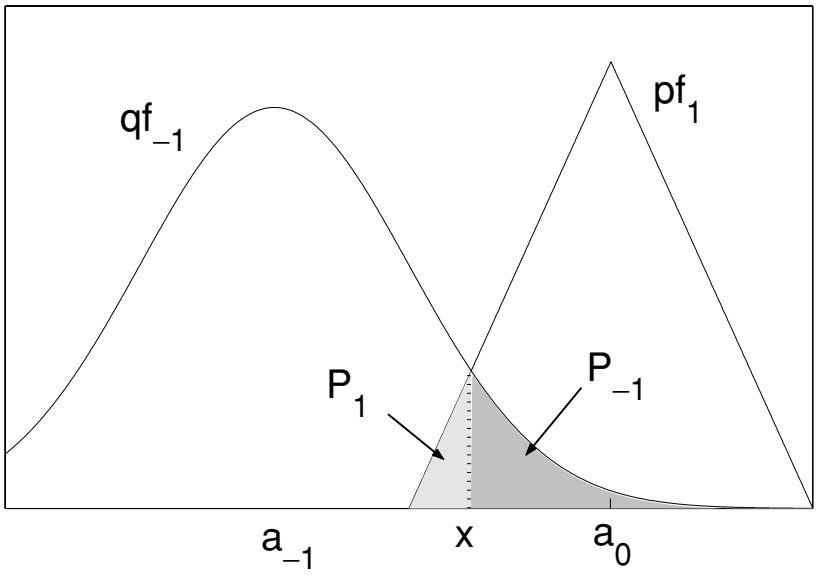

Figure 4: Stoller split problem for two univariate continuous distributions.

When we have separable classes, it is obvious that we should set x^* anywhere in $]b, c[$. The minimum entropy value ($H_E = 0$) also occurs in that interval because $P(E = 0) = 1$. Again we are led to the minimum probability of error.

4 EEM Splits for Mutually Symmetric Distributions

4.1 Critical Points of the Error Entropy. Suppose the two classes $\mathcal{C}_t$, $t \in \{-1, 1\}$ are represented by arbitrary continuous pdf's, $f_t(x)$. We define the center a_t of a distribution as its median. Let us consider, without loss of generality, that class $\mathcal{C}_1$ is centered at 0 and the center of class $\mathcal{C}_{-1}$ lies in the nonpositive part of the real line. Figure 4 depicts this setting.

Theorem 3. *In the univariate two-class problem, the Stoller split x^* is a critical point of the error entropy if the error probabilities of each class at x^* are equal.*

Proof. From formula 2.5, one derives

$$\frac{dH}{dx'} = qf_{-1} \log(P_{-1}) - (qf_{-1} - pf_1) \log(1 - P_{-1} - P_1) - pf_1 \log(P_1). \tag{4.1}$$

A critical point (roots of first derivative) of H must satisfy

$$\frac{dH}{dx'} = 0 \Leftrightarrow \frac{p}{q}\frac{f_1}{f_{-1}} = \frac{\log(P_{-1}/(1-P_{-1}-P_1))}{\log(P_1/(1-P_{-1}-P_1))}. \tag{4.2}$$

If the densities are continuous, the Stoller split x^* is obtained either at a pf_1 versus qf_{-1} intersection $pf_1(x^*) = qf_{-1}(x^*)$ or at $+\infty$ or $-\infty$ (see theorem 1). In the last case, the error probabilities of each class are unequal. In the first case, we have, from equation 4.2,

$$pf_1(x^*) = qf_{-1}(x^*) \Leftrightarrow P_{-1}(x^*) = P_1(x^*), \tag{4.3}$$

where $P_{-1}(x^*)$, $P_1(x^*)$ are the probabilities of error of each class with split point at x^*.

Example: In the uniform example of Figure 2a, the Stoller split can be at any point of $[c, b] = [0.7, 1]$, but the critical point (in this case a maximum) of entropy occurs at the middle point of that interval, which corresponds precisely to the split where the two classes have equal error probabilities.

The above result states the conditions for a correspondence between the Stoller split and an entropy extremum. This means that the EEM principle cannot be applied in general situations. Moreover, theorem 3 says nothing about the nature (maximum or minimum) of the critical point. As we will see, the solution in theorem 3 is not guaranteed to be an entropy minimum. Let us determine the sign of $\frac{d^2H}{dx'^2}\big|_{x^*}$. One has

$$\begin{aligned}\frac{d^2H}{dx'^2} = {} & q\frac{df_{-1}}{dx'}\log\left(\frac{P_{-1}}{1-P_{-1}-P_1}\right) - p\frac{df_1}{dx'}\log\left(\frac{P_1}{1-P_{-1}-P_1}\right) \\ & - \frac{(qf_{-1}-pf_1)^2}{1-P_{-1}-P_1} - \frac{q^2f_{-1}^2}{P_{-1}} - \frac{p^2f_1^2}{P_1}.\end{aligned} \tag{4.4}$$

In order to deal with expression 4.4, we make a simplification by analyzing the case of mutually symmetric distributions defined as:

Definition 1. *Two class distributions represented by probability densities g_1 and g_2 and priors p and q, respectively, are mutually symmetric if $pg_1(a_1 - x) = qg_2(x - a_2)$ where a_t is the center of the density g_t.*

If the classes are mutually symmetric, one must have $p = q = 1/2$ and

$$\left.\frac{df_{-1}}{dx'}\right|_{x^*} = -\left.\frac{df_1}{dx'}\right|_{x^*}. \tag{4.5}$$

In the conditions of theorem 3, we have

$$f_1(x^*) = f_{-1}(x^*) \text{ and } P_{-1}(x^*) = P_1(x^*). \tag{4.6}$$

If we define $\frac{1}{2} f_1(x^*) \equiv f$ and $P_1(x^*) \equiv P$, then

$$\left.\frac{d^2H}{dx'^2}\right|_{x^*} = -2\left(\left.\frac{df}{dx'}\right|_{x^*} \log\frac{P}{1-2P} + \frac{f^2}{P}\right). \tag{4.7}$$

Therefore, for mutually symmetric distributions, we need to analyze only what happens at one side of one of the distribution centers (in this case, a_1). Since we have set $a_1 = 0$, x^* occurs at half distance of the median of $\mathcal{C}_{-1}$ to the origin, somewhere in $]-\infty, 0]$. Let

$$G(x^*) = \frac{df}{dx'}\log\frac{P}{1-2P} + \frac{f^2}{P}, \tag{4.8}$$

where we let fall the dependence of the derivative on x^* in order to simplify notation. $G(x^*)$ plays the key role in the analysis of the error entropy critical points. If the classes are sufficiently distant, that is, $\mathcal{C}_{-1}$ sliding to the left ($x^* \to -\infty$ or x^* tends to the infimum of the support of $\mathcal{C}_1$), then $\frac{df}{dx'} > 0$, and we can rewrite expression 4.8 as

$$G(x^*) = \frac{df}{dx'}\left(\log\frac{P}{1-2P} + \frac{f^2}{\frac{df}{dx'}P}\right). \tag{4.9}$$

Using the results given in section A.1, the second term between the parentheses is finite, while P can be made sufficiently small such that the first term is greater in absolute value than the second one. Thus, $G(x^*) < 0$, and equation 4.7 is positive. Hence, the Stoller split x^* is an entropy minimum.

If the classes are sufficiently close, that is, $\mathcal{C}_{-1}$ sliding rightward ($x^* \to 0$), there are three situations to consider. Define x_M and x_m as the abcissas where f has the mode and the median, respectively. Then:

1. $x_M = x_m$. In this case, f is symmetric, and by the continuity of $G(\cdot)$ and the fact that $G(x_M) > 0$ (since $\left.\frac{df}{dx'}\right|_{x_M} = 0$), $G(x^*)$ is positive in a neighborhood of x_M.

2. $x_M < x_m$. Again, $G(x^*) > 0$ in a neighborhood of x_M, because $G(x_M) > 0$.

3. $x_M > x_m$. We have no guarantee on a sign change in $G(x^*)$.

The first two situations show that $G(x^*)$ changes its sign, which means that the Stoller split turns to be an entropy maximum if the distributions are close enough.

In the third situation we may or may not have a sign change of $G(x^*)$. In fact, as shown in section 4.2.3 for the log normal distribution, we have situations where there is always an entropy minimum in an intersection of the posterior densities, but the Stoller split changes its location as the distributions get closer.

Furthermore, for each probability distribution, the ratio x^*/Δ between the possible solution of $G(x^*) = 0$ and the distribution's scale Δ is a constant. In fact, for two variables X and Y, with $Y = \Delta \cdot X$ (Y is a scaled version of X), we have $x^*_Y/\Delta = x^*_X$.

4.2 Critical Points for Some Distributions. This section presents three examples of univariate split problems that illustrate the results of previous sections. In the first two examples, for the triangular and gaussian distributions, we determine the minimum distance between classes such that the Stoller split is an entropy minimum. We define d/Δ as a normalized distance between the centers of the two classes where $d = a_1 - a_{-1}$ and Δ is the distributions scale. Remember from the end of the previous section that it is only needed to set $\Delta = 1$. We also set $p = q = 1/2$ in all examples. The third example shows that one can have an entropy minimum in an intersection point where the probabilities of error are equal but it is not the location of the Stoller split.

4.2.1 The Triangular Distribution Case. The triangular density function with width (scale) Δ is given by

$$f(x) = \begin{cases} 0, & x < 0 \\ \dfrac{2}{\Delta} - \left(\dfrac{2}{\Delta}\right)^2 \left| x - \dfrac{\Delta}{2} \right|, & 0 \le x \le \Delta \\ 0, & x > \Delta. \end{cases} \tag{4.10}$$

Setting $\Delta = 1$, class $\mathcal{C}_1$ is centered at $1/2$ and class $\mathcal{C}_{-1}$ is moving between $-1/2$ and $1/2$. The Stoller split occurs at $x^* = (1/2 + a_{-1})/2$. Carrying out the computation of $G(x^*)$, one finds that x^* will be a minimum of entropy iff

$$\frac{1}{2} 2^2 \log \frac{\frac{1}{4} 2^2 x^{*2}}{1 - \frac{1}{2} 2^2 x^{*2}} + 2^2 < 0 \Leftrightarrow x^* < \frac{1}{\sqrt{e^2 + 2}} \tag{4.11}$$

and a maximum otherwise.

Thus, for any Δ, the Stoller split is an entropy minimum if

$$\frac{d}{\Delta} > 1 - \frac{2}{\sqrt{e^2+2}} \approx 0.3473. \tag{4.12}$$

4.2.2 The Gaussian Distribution Case. For gaussian distributions, one has $a_t = \mu_t$, where μ_t is the distribution mean of class $\mathcal{C}_t$. $G(x^*)$ can be easily rewritten as a function of d. Indeed, setting $\Delta \equiv \sigma = 1$,

$$G(x^*) \equiv \frac{d}{4\sqrt{2\pi}} \exp(-d^2/8) \log\left(\frac{1-\Phi(d/2)}{2\Phi(d/2)}\right) - \frac{\exp\left(-d^2/4\right)}{4\pi\left(1-\Phi\left(d/2\right)\right)} \tag{4.13}$$

where $\Phi(\cdot)$ is the standard gaussian cumulative distribution function.

If d is below some value, expression 4.13 will be positive, and the Stoller split is an entropy maximum. If it is above, the Stoller split is an entropy minimum. This turning value was numerically determined to be $t_{value} = 1.405231264$.

4.2.3 The Log Normal Distribution Case. The log normal distribution has density

$$g\left(x|\mu,\sigma\right) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right). \tag{4.14}$$

We consider the splitting problem where $f_{-1}(x) \equiv g(x)$ and $f_1(x) \equiv g(-x + a_{-1} + x_m)$ where $x_m \equiv a_1$ is the center (median) of f_1. Note that this is precisely the situation 3 referred to in section 4.1 ($x_M > x_m$). In fact, $x_m = e^{\mu}$ and $x_M = e^{\mu-\sigma^2}$. Figure 5 shows the splitting problem in two different conditions: in Figure 5a, the distributions are distant, and in Figure 5b, the distributions have the inner intersection point at their centers. We found that this intersection is always an entropy minimum (thick solid line), but the Stoller split moves to one of the outer intersections (as we can see from the minimum probability of error curve represented by the dashed line) as the distributions get closer. This illustrates the way theorem 3 was enunciated, because one can have an intersection point with equal probabilities of error and thus an entropy critical point, but it may not correspond to the Stoller split intersection.

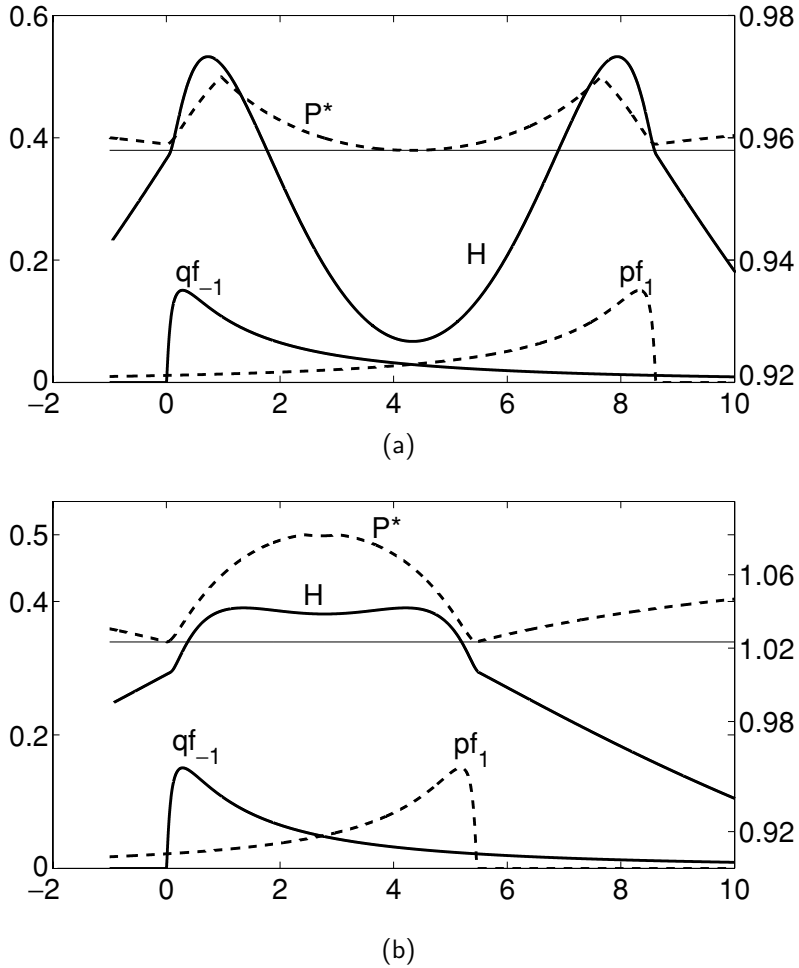

Figure 5: The log normal distribution case. (a) If the distributions are distant, the Stoller split is an entropy minimum at the inner intersection. (b) The inner intersection is still an entropy minimum, but the Stoller split is at one of the outer intersections.

5 EEM Splits in Practice

5.1 The Empirical Stoller Split and MSE. In section 2 we saw how to obtain a theoretical Stoller split for a given problem when the class distributions are known. However, in practice, one has available only a set of examples whose distributions are in general unknown. Stoller (1954)

proposed the following practical rule to choose (x', y') such that the empirical error is minimal:

$$(x', y') = \underset{(x,y)\in\mathbb{R}\times\{-1,1\}}{\arg\min} \frac{1}{N}\sum_{i=1}^{N}\left(I_{\{X_i\le x, T_i\neq y\}} + I_{\{X_i> x, T_i\neq -y\}}\right). \tag{5.1}$$

The probability of error of Stoller's rule converges to the Bayes error for $N\to\infty$ (for details, see Devroye et al., 1996). If we take the MSE cost function,

$$MSE = c\sum_{i=1}^{N}(t_i - y_i)^2, \tag{5.2}$$

where c is a constant,[2] it is easy to see that it is equivalent to Stoller's rule, equation 5.1, in the sense that the same discrimination rule, equation 2.1, is determined. In fact,

$$MSE = c\left[\sum_{X_i\in\mathcal{C}_{-1}}(t_i - y_i)^2 + \sum_{X_i\in\mathcal{C}_1}(t_i - y_i)^2\right] \tag{5.3}$$

$$= c\left[\sum_{X_i\in\mathcal{C}_{-1}} 4I_{\{X_i> x\}} + \sum_{X_i\in\mathcal{C}_1} 4I_{\{X_i\le x\}}\right] \tag{5.4}$$

$$= 4c\sum_{i=1}^{N}\left(I_{\{X_i\le x, T_i\neq -1\}} + I_{\{X_i> x, T_i\neq 1\}}\right), \tag{5.5}$$

which is the same as in equation 5.1 if we take $c = 1/4N$ and use the convention that class $\mathcal{C}_{-1}$ is at the left of the splitting point. Thus, the solution to

$$(x', y') = \underset{(x,y)\in\mathbb{R}\times\{-1,1\}}{\arg\min} MSE \tag{5.6}$$

is the same as in equation 5.1.

5.2 EEM Empirical Procedure. We have to develop a practical rule to minimize (or maximize, depending on the conditions of the problem) the

[2] The value of c (which can be $1/N$ for MSE definition or $1/2$ for derivative simplification reasons) has no influence on the minimization of the cost function.

error entropy,

$$H(x) = -P_{-1}(x)\log P_{-1}(x) - P_1(x)\log P_1(x) - (1 - P_{-1}(x) - P_1(x))\log(1 - P_{-1}(x) - P_1(x)), \quad (5.7)$$

where

$$P_{-1}(x) = \int_x^{-\infty} q f_{-1}(s)ds = q\left[1 - \int_{-\infty}^{x} f_{-1}(s)ds\right] \quad (5.8)$$

$$P_1(x) = p\int_{-\infty}^{x} f_1(s)ds. \quad (5.9)$$

Since we do not know the true class distributions, we estimate them using the gaussian kernel density estimator (Parzen, 1962),

$$f_t(x) \approx \frac{1}{Nh}\sum_{x_i \in \mathcal{C}_t} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - x_i)^2}{2h^2}\right). \quad (5.10)$$

Hence,

$$\int_{-\infty}^{x} f_t(s)ds \approx \frac{1}{N}\sum_{x_i \in \mathcal{C}_t} \Phi\left(\frac{x - x_i}{h}\,\middle|\, 0, 1\right), \quad (5.11)$$

where $\Phi(x|\mu, \sigma^2)$ is the cumulative gaussian distribution, with mean μ and variance σ^2, at x. Expression 5.11 is used to compute and optimize $H(x)$ as in equation 5.7, and expression 5.1 is used to obtain the optimal solution for MSE. The optimization algorithm we have used in our experiments is based on the Golden Section search with parabolic interpolation (Press, Teukolsky, Vetterling, & Flannery, 1992).

5.3 Experiments

5.3.1 Simulated Data: The Two-Class Gaussian Problem. We first studied how the EEM procedure works with simulated gaussian data, where all the conditions can be controlled. To ensure the conditions of theorem 3, two classes with gaussian distribution differing only in location (σ was set to 1) were generated. We also set $p = q = 0.5$. Several experiments were made varying the normalized distance d/σ between classes. Taking into account the t_{value} for gaussian classes, the distance values were chosen so as to have a maximization problem ($d/\sigma = 1$) and two minimization problems ($d/\sigma = 1.5$ and 3), one of them very close to t_{value}. We also varied the number of available training (# *train*) and test (# *test*) patterns for each class. The

Table 1: Test Error (%) and Standard Deviations Obtained with EEM and MSE for the Simulated Gaussian Data.

# *train*	100		1000		100000	
# *test*	EEM	MSE	EEM	MSE	EEM	MSE
$d = 3$; Bayes error: 6.68%						
50	6.79(2.41)	7.02(2.59)	6.75(2.51)	6.72(2.60)	6.75(2.51)	6.66(2.41)
500	6.82(0.83)	7.07(1.00)	6.70(0.81)	6.76(0.81)	6.66(0.81)	6.69(0.81)
5000	6.81(0.30)	7.11(0.59)	6.69(0.25)	6.72(0.27)	6.68(0.25)	6.67(0.25)
50,000	6.81(0.20)	7.11(0.60)	6.69(0.08)	6.74(0.13)	6.68(0.08)	6.68(0.08)
$d = 1.5$; Bayes error: 22.66%						
50	25.23(4.65)	23.22(4.22)	24.67(4.58)	22.74(4.23)	22.61(4.21)	22.54(4.10)
500	25.33(2.75)	23.30(1.54)	24.82(2.48)	22.84(1.36)	22.83(1.38)	22.67(1.32)
5000	25.32(2.49)	23.22(0.84)	24.72(2.15)	22.82(0.48)	22.80(0.46)	22.68(0.42)
50,000	25.46(2.54)	23.27(0.77)	24.83(2.21)	22.81(0.23)	22.82(0.24)	22.67(0.13)
$d = 1$; Bayes error: 30.85%						
50	30.63(4.64)	31.56(4.60)	30.90(4.48)	31.05(4.61)	30.70(4.82)	30.69(4.56)
500	30.95(4.64)	31.56(4.60)	30.88(1.45)	31.01(1.46)	30.80(1.49)	30.75(1.44)
5000	30.93(0.47)	31.37(0.84)	30.87(0.47)	31.04(0.50)	30.84(0.46)	30.84(0.45)
50,000	30.93(0.17)	31.39(0.70)	30.86(0.14)	31.02(0.26)	30.85(0.14)	30.86(0.15)

Notes: Different values of d were used, and the Bayes error was determined for each case. Standard deviations are in parentheses.

solution was determined for both EEM and MSE with the training set and tested with the test set over 1000 repetitions. To determine the value of h to use in each problem, we conducted preliminary experiments where we varied h in order to choose the best one. The final values used were $h = 1.7$, 0.1 and 0.8 for $d = 1$, 1.5, and 3, respectively. As these problems can be solved optimaly, in the Bayes sense, by a unique split, we have determined the Bayes error for each experiment for comparison purposes. Table 1 shows the mean values and standard deviations for the test error of each experiment.

For $d = 1$ and $d = 3$, both EEM and MSE achieve Bayes discrimination if the training sets are asymptotically large, with slightly better results for EEM. However, with small training sets, EEM outperforms MSE. In fact, we encounter less test error and standard deviations for EEM, which means that its solutions have more stability and more generalization capability. Increasing the number of test patterns has the major effect of decreasing the standard deviation of the error estimates.

In this sense, the results for $d = 1.5$ were quite unexpected. As we can see, the results of EEM are always worse than with MSE, mainly for small sample sizes. Further investigation revealed that the problem was due to the proximity of $d = 1.5$ to the turning value. The estimate of entropy has high variance, and the location of extrema is highly dependent on the value

Table 2: Test Error (%) and Standard Deviations Obtained with EEM (Maximization Approach) and MSE for the Simulated Gaussian Data ($d = 1.5$).

# *train*	100		1000		100,000	
# *test*	EEM	MSE	EEM	MSE	EEM	MSE
50	22.95(3.93)	23.22(4.22)	22.78(4.01)	22.74(4.23)	22.47(4.14)	22.54(4.10)
500	22.73(1.28)	23.30(1.54)	22.71(1.32)	22.84(1.36)	22.63(1.33)	22.67(1.32)
5000	22.73(0.41)	23.22(0.84)	22.65(0.43)	22.82(0.48)	22.66(0.41)	22.68(0.42)
50,000	22.75(0.17)	23.27(0.77)	22.67(0.14)	22.81(0.23)	22.67(0.13)	22.67(0.13)

Note: Standard deviations are in parentheses.

of h. To solve this problem, we investigated the possibility of transforming the minimization problem into a maximization problem, getting a more accurate and stable procedure. This is achieved by increasing the value of h (the details are described in section A.2). The performance is increased not only in terms of lower test error but also lower number of iterations needed. Table 2 presents the comparison between MSE and the maximization approach, where h was determined by formula A.5 with $c = 3$. As we can see, EEM now behaves similarly as for $d = 1$ and $d = 3$ above, outperforming the results of MSE.

5.3.2 Real Data. The EEM and MSE procedures were also applied to real data. We've used four data sets: Corkstoppers from Marques de Sá (2001) and Iris, Wine, and Glass from the UCI repository (Newman, Hettich, Blake, & Merz, 1998). We intended to use the previous results for gaussian distributions; therefore, we have conducted hypothesis testing on the normality of the samples and homogeneity of variances. Table 3 shows a brief description of the data used and the results of these tests.

All samples except the ones from Glass verify the normality assumption (for a significance level $\alpha = 0.05$). The homogeneity of variance property can also be ensured for the same significance level, except for Wine and Glass. Thus, we expect a worse performance of EEM in these data sets, because the conditions of theorem 3 are not ensured. Taking into account the d/σ values, we have two minimization (Corkstoppers and Iris petal length) and four maximization problems.

The train and test procedure was a simple holdout method: half of the data set for training and half for testing. This was repeated over 1000 times, varying the train and test sets. The results obtained are shown in Table 4.

The results show that EEM outperforms MSE in most cases with definitely better results in four of the six data sets (according to the $\mu_{EEM} = \mu_{MSE}$ test). Even in Wine, EEM outperformed MSE. In Corkstoppers, the minimization of error entropy performed poorly. This is in agreement with the

Table 3: Description of the Univariate Two-Class Problems Used from Real Data.

	Corkstoppers	Iris			Wine	Glass
x	N	Sepal Length	Sepal Width	Petal Length	Alcalinity of Ash	NA
classes	1 vs. 2		2 vs. 3		1 vs. 2	1 vs. 2
d/σ	1.474	1.036	0.629	2.341	0.717	0.068
Normality	0.97; 0.76	0.58; 0.91	0.45; 0.53	0.25; 0.29	0.18; 0.43	0.04; 0.00
$\sigma_1^2 = \sigma_2^2$ test	0.72	0.15	0.85	0.26	0.01	0.02

Notes: x is the input variable used, and *classes* is the two classes used from each data set. The last two rows show the p-values for the normality and homogeneity of variance tests.

Table 4: Percentage of Test Error for the Univariate Split Problems of Table 3 with EEM and MSE.

	Corkstoppers		Iris		Wine	Glass
EEM	22.94(4.50)	27.25(4.62)	41.25(8.30)	8.15(2.73)	33.64(4.13)	52.64(3.22)
MSE	26.19(4.84)	30.24(5.43)	40.77(8.2)	8.52(3.17)	35.43(4.50)	52.81(3.29)
$\mu_{EEM} = \mu_{MSE}$	0.00	0.00	0.098	0.005	0.00	0.122

Notes: The last row presents the p-values of the test of equality of means $\mu_{EEM} = \mu_{MSE}$. Standard deviations are in parentheses.

results obtained for the gaussian simulated problem with $d = 1.5$. Thus, the results of Corkstoppers in Table 4 were obtained with the maximization procedure using equation A.5 with $c = 3$ to set h.

In Iris petal length, we used the minimization approach with better results than MSE, but it was interesting to notice that the maximization approach achieved even better results: test error of 7.18% and standard deviation 2.78%. We sought an explanation for the difference of the maximization and minimization results and found it on the small number of patterns used each time in the training sets, where each class density is estimated with approximately 25 patterns. Also, the optimal value used for the minimization was a mere $h = 0.16$ (empirically found), which in conjunction with the small number of patterns produces very rough density estimates (see Figure 6a), contrasting with those obtained with a large h (Figure 6b, for the maximization procedure). Furthermore, as the training sets (remember that each experiment is repeated 1000 times) may vary a lot, the same value of $h = 0.16$ for all of them is certainly not an optimal

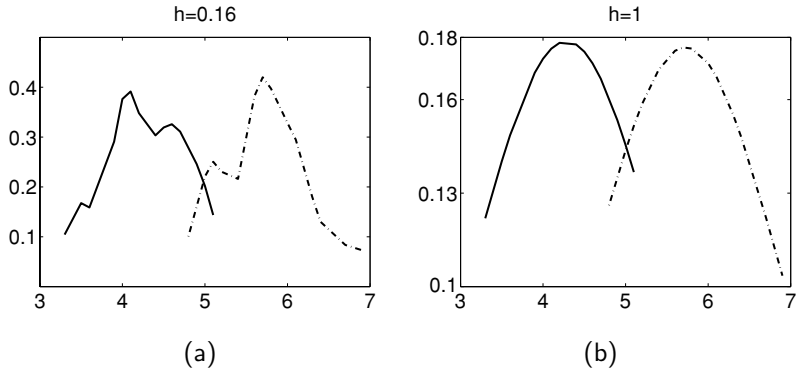

Figure 6: Density estimates of a training set for the two class problem of Iris petal length with $h = 0.16$ in (a), minimization procedure and $h = 1$ in (b), maximization procedure.

choice. On the other hand, as the maximization approach uses larger h, the possible differences between different training sets are smoothed out and have less influence on the final result. This explains the difference between the minimization and maximization results. In conclusion, for very small data sets (when the sample may not be representative of the distribution), one should consider the maximization approach.

6 Discussion and Conclusions

We analyzed the relation between the theoretical Stoller split (univariate two-class discrimination problem) and the error entropy minimization (EEM) principle. Besides the possible practical applications of this analysis to univariate data splitting with EEM (e.g., in tree classifier design, using the popular univariate data splitting approach), the results derived from the analysis are also important as a first step to a needed theoretical EEM assessment when applied to neural networks (e.g., multilayer perceptrons, MLP). For instance, this work has shown that for certain class configurations, one must use entropy maximization instead of minimization.

We started by verifying that for two uniform classes, the EEM principle leads to the optimal classifier for the class of Stoller split decision rules. This optimal solution also corresponds to the optimal decision rule obtained using the minimum probability of error criterion. Thus, Bayes error is also guaranteed in this situation. For general class density functions, it was proven (in theorem 3) that a Stoller split occurs at an entropy extremum only if the error probabilities for both classes are equal; this restricts the applicability of the EEM principle to univariate splitting in the sense that the optimal classifier may not be achieved. Moreover, we showed that for

mutually symmetric distributions and in the conditions of theorem 3, the Stoller split may be either an entropy minimum or maximum, depending on the proximity of the classes. In particular, it was possible to determine the turning proximity values for triangular and gaussian distributions. These were used as a guideline for the empirical procedure, where EEM outperformed MSE, especially for small sample sizes.

With simulated data, we concluded that the EEM principle requires fewer training data than MSE and also fewer iterations of the optimization algorithm. This fast convergence evidence will be studied in more detail in future work, particularly when EEM is applied to general MLP classification.

We also encountered a high sensitivity of the discrimination process to the smoothing parameter, h. This phenomenon has already been reported in previous work (Santos, Alexandre, et al., 2004; Santos, Marques de Sá, et al., 2004; Silva et al. 2005). Meanwhile, our analysis enlightened the fact that in the cases where d/Δ is near the turning proximity value, it is preferable to set h so as to convert the minimization process into a maximization process. Furthermore, our maximization results show that exact density estimation is not needed; a density estimation that is capable of extracting the main characteristics of the data is sufficient. All of these findings are certainly important for future study of the influence of h for MLP classifiers with either threshold or continuous activation functions.

Appendix A: Additional Results

A.1 A Result on the Hölder Exponent

Definition 2. *Let $\alpha \in \mathbb{R}^+$ and $x_0 \in \mathbb{R}$. A function $f : \mathbb{R} \to \mathbb{R}$ is said to be $C^{[\alpha]}(x_0)$ if there exists $L > 0$ and a polynomial P of degree $[\alpha]$[3] such that*

$$\forall \delta > 0 : |x - x_0| < \delta \Rightarrow \left| f(x) - P(x - x_0) \right| \leq L \, |x - x_0|^{\alpha} . \tag{A.1}$$

The maximum value of α that satisfies equation A.1 is known as the Hölder exponent of f at x_0.

The polynomial P is the Taylor expansion of order $[\alpha]$ of f at x_0. The Hölder exponent α measures how irregular f is at the point x_0. The higher the exponent α, the more regular is f. Figure 7 shows the behavior of f in a neighborhood of x_0 for different values of α.

[3] $[\alpha]$ represents the largest integer less than α. If α is not an integer, $[\alpha] \equiv \lfloor \alpha \rfloor$; otherwise, $[\alpha] \equiv \alpha - 1$.

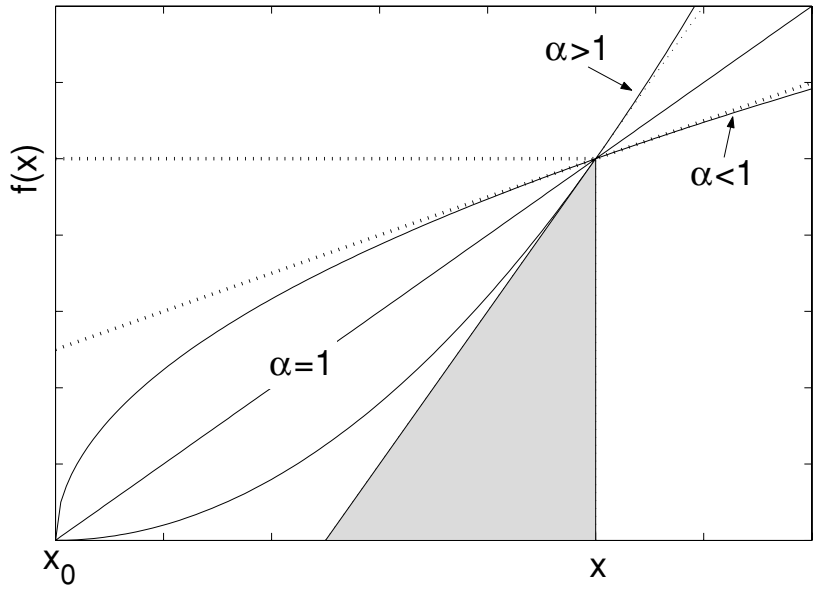

Figure 7: Local behavior of f for different values of α.

Theorem 4. *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function, such that $f \equiv 0$ for $x \leq x_0$ and differentiable for $x > x_0$. If the Hölder's exponent of f at x_0 is α, then*

$$lim_{x \to x_0^+} \frac{f^2(x)}{\left[\int_{x_0}^{x} f(y)dy\right] \frac{df}{dx}(x)} = \frac{\alpha + 1}{\alpha}. \tag{A.2}$$

The idea of the previous theorem is that in a sufficiently small neighborhood of x_0, f behaves like $L(x - x_0)^\alpha$. Then

$$\frac{f^2(x)}{\left[\int_{x_0}^{x} f(y)dy\right] \frac{df}{dx}(x)} = \frac{L^2(x - x_0)^{2\alpha}}{\frac{L(x-x_0)^{\alpha+1}}{\alpha+1} \alpha L(x - x_0)^{\alpha-1}} = \frac{\alpha + 1}{\alpha}.$$

The left-hand side of equation A.2 is also bounded if f has left unlimited support. The proof of this result can be made using a geometrical argument. In fact,

$$\left[\int_{-\infty}^{x} f(y)dy\right] \frac{df}{dx} > \frac{f(x)b}{2} \frac{f(x)}{b},$$

where b is the base of the shadowed triangle in Figure 7.[4] Thus,

$$\frac{f^2(x)}{\left[\int_{-\infty}^{x} f(y)dy\right]\frac{df}{dx}(x)} < 2.$$

A.2 Turning Minimization into Maximization. Estimating a density function with kernel method leads to an estimate with

$$\hat{\mu} = \bar{x}, \qquad \hat{\sigma}^2 = h^2 + s^2 \tag{A.3}$$

where $\bar{x}$ is the sample mean and s^2 is the sample (not corrected) variance. When h is too small, the kernel estimate has a large variance leading to a nonsmooth entropy function. When h is large, we have an oversmoothed density, but entropy is smooth and preserves the extrema. Figure 8 depicts this dichotomy. In Figures 8b and 8c, the values of h are given by the optimal rule for gaussian distributions (Silverman, 1986) and by expression A.5 with $c = 3$, respectively. The vertical solid line shows the theoretical Stoller split for the problem. It is important to note that this is a minimization problem. In practice, the increased h has the effect of approximating classes and thus the maximum instead of the minimum in Figure 8c. This means that it is more efficient to maximize entropy when d/σ is close to the turning value.

How can one set h in order to have a maximization problem? Just ensure that

$$\frac{d}{\sigma} \approx \frac{t_{value}}{c} \tag{A.4}$$

where $c > 1$ and σ is the standard deviation of the estimated density. Thus, with straightforward calculations, one has

$$h^2 \approx \left(\frac{d\ c}{t_{value}}\right)^2 - s^2, \tag{A.5}$$

or h equal to some large value (empirically obtained) if the right-hand side of equation A.5 is nonpositive. An evident choice for c may be $c = t_{value}$, because this implies $d/\sigma = 1$, which is the third gaussian problem of section 5.3.1. The increase in c leads to increased h, and the entropy function becomes smoother. Of course, one cannot increase h indefinitely, because with almost flat H, the optimization algorithm may fail to find its maximum.

[4] Note that the behavior of f in this situation is similar to the case of f with left limited support and $\alpha > 1$.

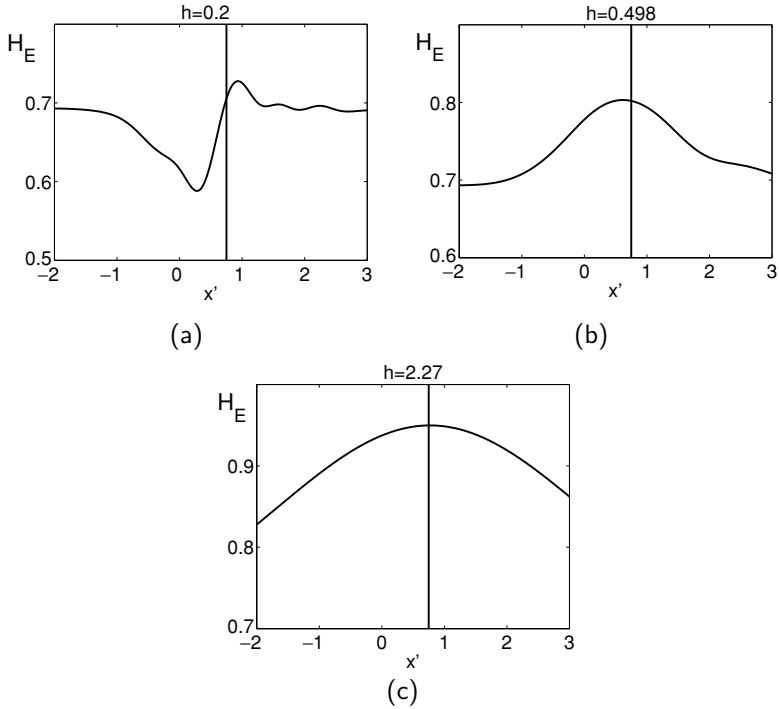

Figure 8: Error entropy for different values of h in the gaussian distribution example with $d = 1.5$.

Appendix B: Proof of Theorem 1

Proof. First, assume that there is no intersection of qf_{-1} with pf_1 (see Figure 9a). Then $P^* = \min(p, q) \leq 1/2$ occurs at $+\infty$ or $-\infty$.

For intersecting posterior densities, one has to distinguish two cases. First, assume that for $\delta > 0$,

$$\begin{aligned} &pf_1(x) < qf_{-1}(x) \quad x \in [x_0 - \delta, x_0] \qquad \text{and} \\ &pf_1(x) > qf_{-1}(x) \quad x \in [x_0, x_0 + \delta], \end{aligned} \tag{B.1}$$

where x_0 is an intersection point (see Figure 9b). The probabilities of error at x_0 and $x_0 - \delta$ are

$$P(x_0) = p\left[\int_{-\infty}^{x_0-\delta} f_1(t)dt + \int_{x_0-\delta}^{x_0} f_1(t)dt\right] + q\int_{x_0}^{+\infty} f_{-1}(t)dt \tag{B.2}$$

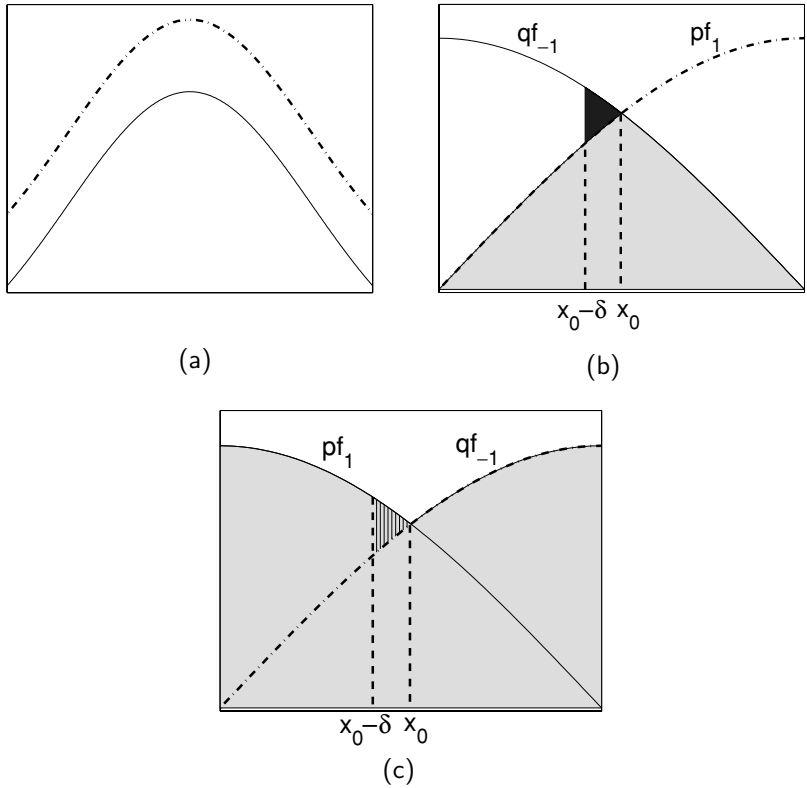

Figure 9: Possible no-intersection or intersection situations in a two-class problem with continuous class-conditional density functions. The light shadowed area in b and c represents $P(x_0)$ where x_0 is the intersection point. The dark shadowed area in b represents the amount of error probability added to $P(x_0)$ when the splitting point is deviated to $x_0 - \delta$. The dashed area in c is the amount of error probability subtracted from $P(x_0)$ when the splitting point is deviated to $x_0 - \delta$.

$$P(x_0 - \delta) = p \int_{-\infty}^{x_0-\delta} f_1(t)dt + q \left[\int_{x_0-\delta}^{x_0} f_{-1}(t)dt + \int_{x_0}^{+\infty} f_{-1}(t)dt \right]. \qquad \text{(B.3)}$$

Hence,

$$P(x_0) - P(x_0 - \delta) = p \int_{x_0-\delta}^{x_0} f_1(t)dt - q \int_{x_0-\delta}^{x_0} f_{-1}(t)dt < 0 \qquad \text{(B.4)}$$

by condition B.1. It is easily seen, using similar arguments, that $P(x_0) - P(x_0 + \delta) < 0$. Thus, x_0 is a minimum of $P(x)$. Now, suppose that (see Figure 9c)

$$\begin{aligned} &pf_1(x) > qf_{-1}(x) \quad x \in [x_0 - \delta, x_0] \qquad \text{and} \\ &pf_1(x) < qf_{-1}(x) \quad x \in [x_0, x_0 + \delta]. \end{aligned} \tag{B.5}$$

Then x_0 is a maximum of $P(x)$. This can be proven as above or just by noticing that this situation is precisely the same as above but with a relabeling of the classes. For relabeled classes, the probability of error $P^{(r)}(x)$ is given by

$$\begin{aligned} P^{(r)}(x) &= p(1 - F_{-1}^{(r)}(x)) + q F_1^{(r)}(x) \\ &= 1 - [q(1 - F_{-1}(x)) + pF_1(x)] = 1 - P(x). \end{aligned} \tag{B.6}$$

Thus, $P^{(r)}(x)$ is just a reflection of $P(x)$ around $1/2$, which means that $P(x)$ maxima are $P^{(r)}(x)$ minima and vice versa. The Stoller split is chosen as the minimum up to a relabel (see expression 2.3).

Acknowledgments

This work was supported by the Portuguese FCT-Fundação para a Ciência e a Tecnologia (project POSC/EIA/56918/2004). L.M.S. is also supported by FCT grant SFRH/BD/16916/2004.

References

Devroye, L., Gyorfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition.* Berlin: Springer-Verlag.

Erdogmus, D., & Principe, J. (2000). Comparison of entropy and mean square error criteria in adaptive system training using higher order statistics. In *Proceedings of the Intl. Conf. on ICA and Signal Separation* (pp. 75–80). Helsinki, Finland.

Erdogmus, D., & Principe, J. C. (2002). An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems. *IEEE Transactions on Signal Processing, 50*(7), 1780–1786.

Kapur, J. (1993). *Maximum-entropy models in science and engineering* (rev. ed). New York: Wiley.

Kullback, S. (1959). *Statistics and information theory.* New York: Wiley.

Linsker, R. (1988). Self-organization in a perceptual network. *IEEE Computer, 21*, 105–117.

Marques de Sá, J. (2001). *Pattern recognition: Concepts, methods and applications.* Berlin: Springer-Verlag.

Newman, D., Hettich, S., Blake, C., & Merz, C. (1998). *UCI repository of machine learning databases.* Irvine: University of California, Irvine, Department of

Information and Computer Sciences. Available online at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics, 33*, 1065–1076.

Press, W., Teukolsky, S., Vetterling, W., & Flannery, B. (1992). *Numerical recipes in C: The art of scientific computing*. Cambridge: Cambridge University Press.

Principe, J. C., Xu, D., & Fisher, J. (2000). Information theoretic learning. In S. Haykin (Ed.), *Unsupervised adaptive filtering, Vol. 1: Blind source separation* (pp. 265–319). New York: Wiley.

Santos, J., Alexandre, L., & Marques de Sá, J. (2004). The error entropy minimization algorithm for neural network classification. In *Int. Conf. on Recent Advances in Soft Computing*. Nottingham, U.K.: Nottingham Trent University.

Santos, J., Marques de Sá, J., Alexandre, L., & Sereno, F. (2004). Optimization of the error entropy minimization algorithm for neural network classification. In C. Dagli, A. Buczak, D. Enke, M. Embrechts, & O. Ersoy (Eds.), *Intelligent engineering systems through artificial neural networks* (Vol. 14, pp. 81–86). New York: American Society of Mechanical Engineers.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423, 623–656.

Silva, L., Marques de Sá, J., & Alexandre, L. (2005). Neural network classification using Shannon's entropy. In *European Symposium on Artificial Neural Networks*. Bruges, Belgium: d-side publications.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Stoller, D. (1954). Univariate two-population distribution free discrimination. *Journal of the American Statistical Association, 49*, 770–777.

Received February 2, 2005; accepted February 17, 2006.