Neural Network Interest Group

Título/*Title*:
Data Classification with the Linear Perceptron and the EEM Principle II

Autor(es)/*Author(s)*:
Luís M. Silva

Relatório Técnico/*Technical Report* No. 4 /2007

FEUP/INEB, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal

Título/*Title*:

Data Classification with the Linear Perceptron and the EEM Principle II

Autor(es)/*Author(s)*:

Luís M. Silva

Relatório Técnico/*Technical Report* No. 4 /2007

Publicado por/*Published by*: NNIG. http://paginas.fe.up.pt/~nnig/

Contents

Chapter 1

Introduction

Information Theoretic Learning (ITL) is an area of research enjoying a growing interest and promising new and important breakthroughs in many applications. The introduction of ITL can be traced back at least to [10] who introduced the maximization of mutual information between the input and output of a neural network (the *infomax* principle) as an unsupervised method that can be applied, for instance, for feature extraction. But the real blossom of ITL dates back from more recent years when the minimization of Rényi's quadratic entropy of data errors for solving regression problems was proposed [4]. This was followed by a wide panoply of ITL theoretical results and applications developed by Príncipe and co-workers, namely in time series prediction [5], feature extraction, clustering [6, 9] and blind source separation [8, 3]. The rationale is as follows. Having an adaptive system with output variable Y and target variable T, the *minimization of the error entropy* (MEE), that is the entropy of $E = T - Y$, implies a reduction of the expected information contained in the error, leading to the maximization of the mutual information between the desired target and the system output [4, 5]. This means that the system is learning the target variable. Entropy-based cost functions, since they depend on the whole probability distribution of E, reflect the global behavior of the error distribution; therefore, learning systems with entropic cost functions can be expected to often outperform those using the classic and popular mean square error (MSE) cost, which only reflects the second-order statistics of the error. The outperformance of MEE over MSE has been shown in the cited works, a good example of which is the prediction of the Mackey-Glass temporal series described in [5]. An objection that could be raised to using the MEE principle is the need to estimate the probability density function (pdf) of E, in the case of continuous

error distributions. Now, it is a well-known fact that accurate pdf estimation may be a tougher problem than having to solve a related regression or classification problem. However, it turns out that when applying the MEE principle using Rényi's entropy, pdf estimation is short-circuited altogether [11]. Even if one uses Shannon's entropy usually a simple and coarse pdf estimate is all that is needed [18].
The application of the MEE principle to solving data classification problems has been carried out by our team and divulged in several papers, either using MLP's [15, 13, 18, 16] or recurrent networks [1] (the principle is coined EEM in these references.). It has been applied with success in classifiers using a kernel-based approach [7]. We have also applied entropic cost functions with excellent results in a new data clustering algorithm [14]. Despite the evidence of good performance provided by the experimental results presented in these references, very little is known about the theoretical properties of MEE when applied to data classification. Let us consider a classification problem with a set of classes $\Omega = \{\omega\}$ and a parametric machine (parameter set $W = \{w\}$) performing a mapping $Y = \varphi_w(X)$ where X and Y are the input and output spaces, respectively. The machine is trained by some algorithm in order to minimize a risk functional on the parameter set W of the function class $\Phi = \{\varphi_w\}$, implemented by the classifier, which is often written for continuous data distributions as

$$\min_W R_\Phi = \min_W \sum_\Omega P(\omega) \int_{X,T} L(t,y) dF(t,x|\omega) \quad \text{with } y = \varphi_w(x)$$

where T is the target space, $F(t,x|\omega) \equiv F_{T,X}(t,x|\omega)$ is the joint cumulative distribution and the $P(\omega)$ are prior probabilities. The target-output distance, i.e. the cost function $L(\cdot)$, can be chosen in various ways. For instance, for MSE, $L(t,y) = (t-y)^2$ and for cross-entropy and two-class problems with $Y \in [0,1]$ and $T \in \{0,1\}$, $L(t,y) = t \ln y + (1-t)\ln(1-y)$. Minkowski and exponentially weighted distances have also been proposed. The risk functional for MEE is written not as a distance functional but instead as a functional of the error $E = T - Y$ pdf $f(e) \equiv f_E(e)$(assuming it exists), namely as -$\int_E \ln f(e|\omega) dF(e|\omega)$ for the Shannon entropy of the error, or as $\frac{1}{1-\alpha} \ln \int_E f(e|\omega)^{\alpha-1} dF(e|\omega)$ for the Rényi entropy. Thus the MEE functional reflects the whole error pdf, whereas the popular MSE functional only reflects the error variance. The main problem in data classification called from now on the *classifier problem* is the possibility of attaining the minimum probability of error afforded by the machine architecture, that is, by the family of functions Φ, for some w^*, the so-called *optimal solution*. Let us denote the

minimum probability of error, achievable in Φ by $\min_W Pe_\Phi$[1]. From now on whenever we talk of optimal solution, w^*, we always mean optimal in the $\min_W Pe_\Phi$ sense. The classifier problem corresponds to the following question: does $\min_W R_\Phi$ imply $\min_W Pe_\Phi$? (Note that $\min_W Pe_\Phi$ corresponds in the distance functional to setting $L(t,y) = \{0, \text{ if } t = y\ ; 1, \text{ otherwise}\}$; however we are only interested in risk functionals with continuous integrands, for which efficient optimization algorithms exist.) For instance, if hypothetically $\min_W R_\Phi$ does not lead to $\min_W Pe_\Phi$, one has to conclude that a risk functional is being used which fails to adequately take into account the whole Φ set complexity. One should then turn to another risk functional. This essential problem has been somewhat overlooked in past literature. Concerning MSE, the main and often mentioned results are that for Gaussian distributions MSE yields the optimal regression solution and that the outputs of a neural network (NN) trained with MSE correspond to Bayesian posterior probabilities [2, 12], which allow some confidence that MSE will also perform well in classification problems. However, MSE may fail for some families of error pdf's where MEE performs in the optimal way, as shown in Appendix .1. Another type of problem where MSE will completely fail is when the error data is characterized by a fat-tail distribution, such as sometimes encountered in financial time series. Take the Cauchy distribution. Empirical variances computed in a Cauchy time series vary erratically, since the Cauchy distribution has no variance; however, it does have a finite Shannon entropy; so the application of MEE to such time series is not a problem. Since MEE is a more sophisticated approach than often used MSE or CE, which takes into account the whole distribution of the errors, and given the large amount of good experimental results obtained with MEE, it really seems worth to investigate the classifier problem with MEE. Along this investigation many interesting aspects and new insights come to light. In a previous work [17] we have showed that for univariate data and the Stoller split setting [20] (a popular setting in decision trees) the MEE principle does not always lead to $\min_W Pe_\Phi$ and we were able to rigorously state the very general conditions when it does. In the present work we go several steps further. We investigate the behavior of a simple perceptron trained with MEE using both discrete and continuous activation functions (a.f.). Although the analysis is carried out in two-class simple perceptrons, what really matters is the MEE behavior in realistic situations of univariate families of error pdf's. The main conclusions can therefore be extrapolated to

[1]For some architectures $\min_W Pe_\Phi$ may correspond to the optimal Bayes error. However, this issue will not occupy us here.

more complex classifiers.
The organization of the paper is as follows. In the following section 2 we introduce notation and present the error entropy expressions for both discrete and continuous cases. We also show in section 2 that for data classification the MEE principle is harder to apply than for regression. In Section 3 we analyze the classifier problem for a simple perceptron with a threshold a.f.. The error distribution is therefore discrete and this analysis completes our previous work [17] extrapolating some previous findings and demonstrating some interesting and even surprising results. In Section 4 we analyze the classifier problem for a simple perceptron with sigmoidal a.f. and explain why the MEE principle works in practice besides some negative theoretical findings. Finally, in section 5 we draw the main conclusions.

1.1 MEE is Harder for Classification than for Regression

1.1.1 The Error Entropy for Data Classification

We consider two-class problems where a given instance $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$ from X is to be classified in one of two classes, $\mathcal{C}_{\text{-}1}$ or $\mathcal{C}_1$, the target set is $T \in \{-1, 1\}$, and a machine (e.g., NN) implements a parameterized family $\Phi = \{\varphi_w\}$, $w \in W$, and issues a single output $y \in [-1, 1]$. Any other supports for T and Y could be used. The ones indicated make computations easier. The random variable (r.v.) Y may be discrete (e.g., as a result of the NN having threshold functions as a.f.) or continuous (e.g., as the result of sigmoidal a.f.). The distribution of the error r.v., $E = T - Y$, for two-class problems, is given by (see Appendix .2):

$$\text{Continuous case:} \quad f_E(e) = \sum_{t \in T} \pi_t f_{Y|t}(t - e) \quad e \in [-2, 2]; \tag{1.1}$$

$$\text{Discrete case:} \quad p_E(e) = \sum_{t \in T} \pi_t p_{Y|t}(t - e)\delta(e, t - y) \quad e \in \{-2, 0, 2\}; \tag{1.2}$$

with priors $\pi_t = P(T = t)$, also denote p, q for π_1, $\pi_{\text{-}1}$, respectively. We also denote by $f_{Y|t}$ the conditional pdf $f_Y(.|t)$. Likewise for $p_{Y|t}$. For the continuous case, the error Shannon's entropy $H_S(E)$ can then be decomposed as:

$$H_S(E) = pH_{S|1} + qH_{S|\text{-}1} + H_S(T) \tag{1.3}$$

where $H_{S|t}$ is the error Shannon's entropy[2] for class $\mathcal{C}_t$ and $H_S(T) = \sum_{t \in T} \pi_t \ln \pi_t$ is the prior Shannon's entropy. This formula is the consequence of disjoint integration supports. For the discrete case, and defining $P_{t \in \{-1,1\}}$ as the probability of error for class $\mathcal{C}_t$, the entropy formula becomes

$$H_S = -\left[P_{\text{-}1} \ln(P_{\text{-}1}) + P_1 \ln(P_1) + (1 - P_{\text{-}1} - P_1) \ln(1 - P_{\text{-}1} - P_1)\right]. \quad (1.4)$$

The distributions and entropies are functions of w, the machine parameter vector, although we omit this dependency for the sake of simpler notation.

1.1.2 The Minimum of the Error Entropy

Looking at (1.3) and since $H_S(T)$ is a constant, $\min H_S = \min\{pH_{S|1} + qH_{S|\text{-}1}\}$. Thus, in general one can say nothing about the minimum (location and value) since it will depend on the particular shapes of $H_{S|t}$ as functions of w, and the particular value of p. However, with iso-entropic distributions, i.e. whenever $H_{S|1}(w) = H_{S|\text{-}1}(w)$, $\forall w$, one only has to study one of the $H_{S|t}$. The situation is even more complicated for the discrete case where no decomposition in sub-entropies is possible.

1.1.3 The Minimum of the KL Divergence

An important result concerning the error (Shannon's) entropy minimum was shown by [5]. These authors demonstrated that the MEE principle corresponds to the minimum of the Kullback-Leibler (KL) divergence. This "probability density matching" result was demonstrated for the regression setting. However, for the data-classification setting two difficulties arise:

1. For the regression setting one may write $f_E(e) = f_{Y|x}(d-e|x)$ as in the cited paper, since there is only *one* distribution of y values and d can be seen as the mean of the y values. However, for the classification setting one has to write: $f_{E|t}(e|t) = f_{E|t,x}(d-e|t,x)$. That is, one has to study what happens to each class conditional distribution, individually; and, therefore, to individually study the KL divergence relative to each class

[2]From now on we will omit the dependency on E and simply denote $H_S(E) \equiv H_S$.

conditional distribution, that is:

$$\text{Continuous case:} \qquad KL_t = \int\int f_{XY|t}(x,y) \ln \frac{f_{XY|t}(x,y)}{d_{XY|t}(x,y)} \quad (1.5)$$

$$\text{Discrete case:} \qquad KL_t = \sum_X \sum_y P_{XY|t}(x,y) \ln \frac{P_{XY|t}(x,y)}{D_{XY|t}(x,y)} \quad (1.6)$$

where $d_{XY|t}(x,y)$ (or $D_{XY|t}(x,y)$) is the desired input-output probability density (or mass) function.

2. The KL divergence is *undefined* whenever $d_{XY|t}(x,y)$ (or $D_{XY|t}(x,y)$) has zeros in the supports of X and Y. This problem, which may or may not be present in the regression setting, is always present in the classification setting, since the *desired* input-output probability density or mass functions are continuous and discrete Dirac functions, respectively.

Even if we relax the conditions on the desired input-output probability density or mass functions, for instance by choosing functions with no zeros on the Y support but sufficiently close to Dirac functions, we may not yet reach the MEE condition for classification because of section 1.1.2: attaining the KL minimum for a class conditional distribution, says nothing about the other class conditional distribution and about H_S.

Chapter 2

Perceptron with Discrete Errors

2.1 The General Setting

The perceptron with threshold a.f. implements a linear discriminant

$$y = \begin{cases} 1, & \sum_{i=1}^{d} w_i x_i + w_0 \geq 0 \\ -1, & \sum_{i=1}^{d} w_i x_i + w_0 < 0 \end{cases}, \tag{2.1}$$

where $w_i,\ i = 0, \ldots, d$ are real parameters. Geometrically, the decision surface, given by $\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$ (where $\mathbf{w}^{\mathrm{T}} = (w_1, \ldots, w_d)$) is an hyperplane. We now study the data classification problem in light of the MEE principle. More precisely, we question if the solution (hyperplane) obtained by minimizing the error entropy corresponds to the optimal solution. The output Y in (2.1) and the target T (class membership) are discrete random variables. The error r.v. $E = T - Y$ takes value in $\{-2, 0, 2\}$ with probabilities $P(E = -2) = P_{\text{-}1}$, the probability of misclassifying a $\mathcal{C}_{\text{-}1}$ pattern, $P(E = 2) = P_1$, the probability of misclassifying a $\mathcal{C}_1$ pattern and $P(E = 0) = 1 - P_{\text{-}1} - P_1$, the probability of making a correct classification. The error entropy is given by formula (1.4). For the decision rule corresponding to (2.1) we compute

$$P_{\text{-}1} = P(Y = 1, T = -1) = P(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 \geq 0, T = -1) = q(1 - F_{z|\text{-}1}(0)), \tag{2.2}$$

$$P_1 = P(Y = -1, T = 1) = P(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 \leq 0, T = 1) = pF_{z|1}(0), \tag{2.3}$$

where $F_{z|t}(0) = P(z \leq 0 | T = t)$ is the conditional distribution value at the origin of the univariate r.v. $z = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$. We now consider two cases:

1. Classes with univariate distribution. Here, $d = 1$ and we may write

$$wx + w_0 \geq 0 \iff x \geq -\frac{w_0}{w}. \tag{2.4}$$

Surely, $w \neq 0$. Without loss of generality, we assume class $\mathcal{C}_1$ at the right of class $\mathcal{C}_{-1}$. Hence, we may also consider $w = 1$ and the decision rule becomes

$$x \quad \text{belongs to} \quad \mathcal{C}_1 \quad \text{if} \quad x \geq -w_0$$

This is the Stoller split setting already studied [17].

2. Classes with bivariate distribution. In this case, $d = 2$, and we write

$$\sum_{i=1}^{2} w_i x_i + w_0 \geq 0 \iff w_1 x_1 + w_2 x_2 + w_0 \geq 0,$$

where at least one of w_1 or w_2 must be non-zero. Three situations can occur:

   (a) $w_1 = 0$ and $w_2 \neq 0$.
   The decision surface is the horizontal line given by $x_2 = -\frac{w_0}{w_2}$

   (b) $w_1 \neq 0$ and $w_2 = 0$.
   The decision surface is the vertical line given by $x_1 = -\frac{w_0}{w_1}$

   (c) $w_1, w_2 \neq 0$.
   The decision surface is the general line given by

$$x_2 = -\left(\frac{w_1}{w_2} x_1 + \frac{w_0}{w_2}\right). \tag{2.5}$$

   Note that cases (a) and (b) are quite similar to (2.4), the Stoller split problem, but as we will see later they are not completely similar.

We now proceed to considering Gaussian distributions for the classes.

2.2 The Case of Two Gaussian Classes

We consider the two-class problem with input data having bivariate Gaussian distributions. From the previous discussion, we see that it is crucial to determine the distribution of $z = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$. For that purpose, we take into account that Gaussianity is preserved under linear transformations:

Property 1. *If* $\mathbf{x} = (x_1, \ldots, x_d)^T$ *has multivariate Gaussian distribution with mean vector* $\boldsymbol{\mu}$ *and covariance matrix* Σ*, i.e* $\mathbf{x} \sim G_d(\boldsymbol{\mu}, \Sigma)$*,* $\mathbf{w}_0 \in \mathbb{R}^m$ *and* $\mathbf{W}$ *is a* $m \times d$ *real matrix, then:*

$$\mathbf{z} = \mathbf{W}\mathbf{x} + \mathbf{w}_0 \sim G_m(\mathbf{W}\boldsymbol{\mu} + \mathbf{w}_0, \mathbf{W}\Sigma\mathbf{W}^T).$$

We now consider two classes such that

$$\mathcal{C}_{t\in\{\text{-}1,1\}},\ :\ \mathbf{x} \sim G_2(\boldsymbol{\mu}_t, \Sigma_t) \Rightarrow z \sim G_1(\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}_t + w_0, \mathbf{w}^{\mathrm{T}}\Sigma_t\mathbf{w}).$$

Hence, for $t \in \{-1, 1\}$, we have

$$F_{z|t}(0) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\sqrt{\mathbf{w}^{\mathrm{T}}\Sigma_t\mathbf{w}}} \exp\left(-\frac{(x - \mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}_t - w_0)^2}{2\mathbf{w}^{\mathrm{T}}\Sigma_t\mathbf{w}}\right) dx \tag{2.6}$$

$$= \Phi\left(-\frac{\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}_t + w_0}{\sqrt{\mathbf{w}^{\mathrm{T}}\Sigma_t\mathbf{w}}}\right) \tag{2.7}$$

and therefore

$$P_{\text{-}1} = \pi_{\text{-}1}\left(1 - \Phi\left(-\frac{\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}_{\text{-}1} + w_0}{\sqrt{\mathbf{w}^{\mathrm{T}}\Sigma_{\text{-}1}\mathbf{w}}}\right)\right), \tag{2.8}$$

$$P_1 = \pi_1\ \Phi\left(-\frac{\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}_1 + w_0}{\sqrt{\mathbf{w}^{\mathrm{T}}\Sigma_1\mathbf{w}}}\right). \tag{2.9}$$

To further investigate this two-class problem in light of the MEE principle we assume, without loss of generality, the following:

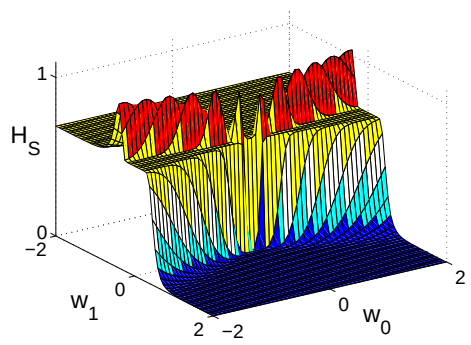
1. Considering $\boldsymbol{\mu}_t = (\mu_{t1}, \mu_{t2})$ for $t \in \{-1, 1\}$, we set $\mu_{t2} = 0$ and $\mu_{\text{-}11} = -\mu_{11}$ with $\mu_{11} > 0$; i.e, the centers of the classes lie in the horizontal axis and are symmetric to each other. Notice that every possible class configuration can be reduced to this case by applying shifts and rotations. As this does not alter the probabilities $P_{\text{-}1}$ and P_1, H_S is only shifted and rotated, preserving the extrema.

2. $\Sigma_{\text{-}1} = \Sigma_1 = I$. By assuming equal covariances, the optimal decision surface is linear (a line in this case). Also, assuming the identity matrix for the covariances corresponds to spherical distributions and allows important simplifications in the above formulas.

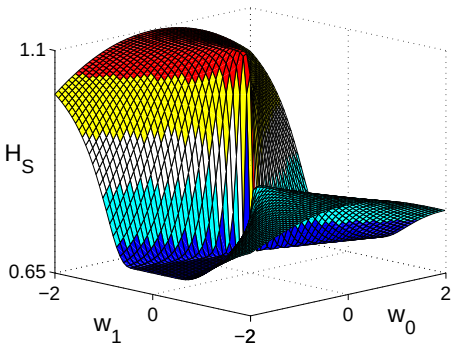
With these assumptions, it is easy to see that the optimal solution $\mathbf{w}^* = (w_1^*, w_2^*, w_0^*)^{\mathrm{T}}$ corresponds to the vertical line $x_1 = 0$ and the optimal decision is to classify $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$ in $\mathcal{C}_1$ if $x_1 \geq 0$. This means that $w_0^* = w_2^* = 0$ and w_1^* must be a positive real number (to give the correct orientation of the classes).

2.2.1 Graphical Analysis

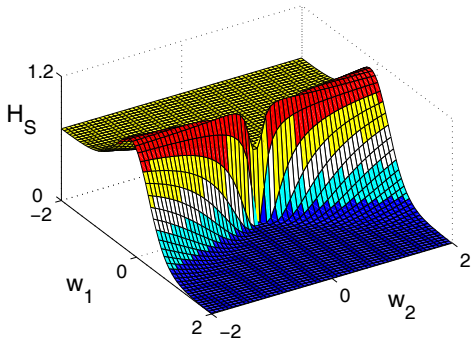
Due to representational reasons, we must fix one of the parameters w_1, w_2 or w_0. As we have some prior knowledge about the solutions we start by setting $w_2 = 0$ and plot H_S as a function of w_1 and w_0 in Figures 2.1(a) and 2.1(b).



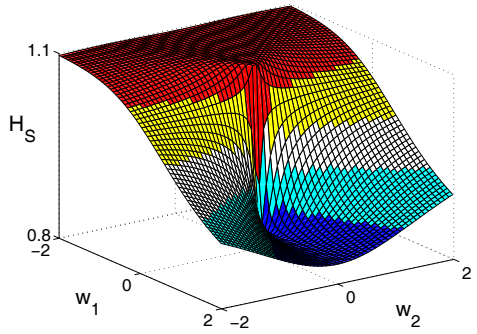
(a) $\mu_{11} = 5$



(b) $\mu_{11} = 0.5$



(c) $\mu_{11} = 5$



(d) $\mu_{11} = 0.5$

Figure 2.1: H_S for different values of $\mu_{11} = -\mu_{\text{-}11}$. From left to right we decrease the distance between the classes. Also, the top figures were drawn with $w_2 = 0$, while the bottom ones were drawn with $w_0 = 0$.

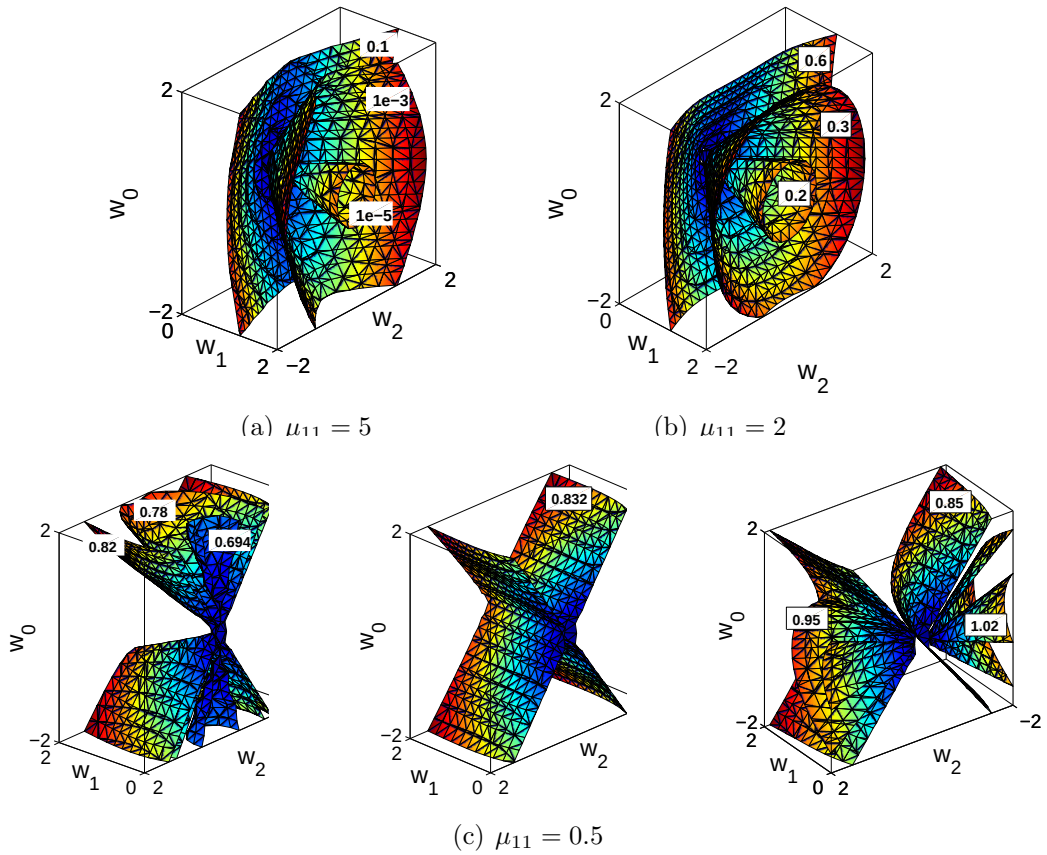
Note that we are assuming a vertical line with freedom to make shifts as the solution to the problem. This is in fact equivalent to the Stoller split case. We may distinguish two regions defined by $w_1 > 0$ and $w_1 < 0$.

- $w_1 > 0$
  When the classes are distant, the optimal solution is obtained at $w_0 = 0$, although small shifts of the line are also acceptable (the flat region

in Figure 2.1(a)). In fact, there are infinite near-optimal solutions with approximately the same entropy ($H_S \approx 0$). This is because the probabilities P_1 and $P_{\text{-}1}$ are not greatly affected by (small) shifting when the classes are distant. However, when the classes get extremely close (Figure 2.1(b)), we obtain a *local* maximum of the entropy for $w_0 = 0$, which is in accordance to the results obtained for Stoller splits [17].

- $w_1 < 0$
  In this case, a swapped classification is being performed, $\mathcal{C}_{\text{-}1} \leftrightarrow \mathcal{C}_1$. The same behaviors as for $w_1 > 0$ are observed.

If we now set $w_0 = 0$, we are considering a solution given by a line passing through the origin but capable of rotating. Let us analyze the $w_1 > 0$ case, by inspecting Figures 2.1(c) and 2.1(d). As expected, the minimum of H_S is attained when $w_2 = 0$, but now it will not turn into a maximum when the classes get closer. Simply, the flat region disappears, because decision boundaries with slope are less tolerable here (there is more probability of error). Thus, we encounter different behavior for $w_2 = 0$ and for $w_0 = 0$. A natural question then arises: what is the behavior of H_S when all the parameters are free to vary? More precisely, when training a learning machine that implements a hyperplane as the decision surface (like the single perceptron), there is, in general, no prior information that one or more of the parameters w_1, w_2 or w_0 should be set to zero (assuming appropriate data shift and rotation). Does the optimal set of parameters also correspond to an entropy minimum? Does it turn to a maximum when the classes get closer (as in the Stoller split case)? We start investigating these questions by inspecting the surface levels of H_S, the equivalent to contour levels in the two variable case. In other words, we examine the surfaces $H_S(w_1, w_2, w_0) = c$ for increasing or decreasing values of $c \in \mathbb{R}$. Figure 2.2 shows some surface levels (iso-entropy surfaces or iso-entropics) of H_S. For distant classes, Figures 2.2(a) and 2.2(b) show that as one decreases the value of c, the iso-entropics converge to the positive w_1 axis, meaning that $H_S(w_1, 0, 0)$ for $w_1 > 0$ has the lowest entropy value. On the other hand, when the classes get closer the behavior is completely different. This case is split into three subfigures in Figure 2.2(c), where from left to right, we gradually increase the value of c. We find that when c is decreased to its minimum, the iso-entropics converge to the w_0 axis (left figure). On the other hand, when c is increased to its maximum (for $w_1 > 0$), the iso-entropics converge to the axis w_2 (right figure). The positive w_1 axis appears for an intermediate value of c, shown



(a) $\mu_{11} = 5$ (b) $\mu_{11} = 2$

(c) $\mu_{11} = 0.5$

Figure 2.2: Surface levels of H_S (values of c also shown). Figure (c) is split into three subfigures with increasing value of c from left to right.

in the middle figure of Figure 2.2(c), which means that this solution (in fact, the optimal solution) is not a *global* minimum nor maximum of the entropy. In fact, as we found with Figure 2.1(b) the positive w_1 axis is a *local* maximum.

2.2.2 First and Second Order Information

Despite the above graphical suggestions one cannot conclude with confidence the exact nature of the solutions. We now study their behavior from an analytical point of view, using first and second order information about H_S (first and second order derivatives). In what follows we omit several expressions due to their complexity and length. It is easy to show that for the

class configuration stated in section 2.2 (the centers of the classes lie in the horizontal axis x_1 and are symmetric to each other) we have $H_S(w_1, 0, 0) = c_1 \ \forall w_1 > 0$ and $H(w_1, 0, 0) = c_2 \ \forall w1 < 0$, where $c_1, c_2 \in \mathbb{R}$. Moreover, it can be proved that $c_1 \to 0$ (100% correct classification) and $c_2 \to \ln(0.5)$ (ditto, with swapped class labels), as we increase the distance between the classes. Computing the gradient of H_S we find that vectors of the form $\bar{\mathbf{w}} = (w_1, 0, 0)^{\mathrm{T}}$ for $w_1 \neq 0$ are critical points of H_S or, in other words, that $\nabla H_S(\bar{\mathbf{w}}) = \mathbf{0}$. The nature of these critical points can be further investigated using second order information about H_S, given by its Hessian matrix. We restrict our study to the following cases:

1. $\mu_{11} = 5$ and $w_1 > 0$
   The Hessian matrix $\nabla^2 H_S$ at $\bar{\mathbf{w}}$ is given by

$$\nabla^2 H_S(\bar{\mathbf{w}}) \approx \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{0.4809}{w_1^2} & 0 \\ 0 & 0 & \frac{0.3527}{w_1^2} \end{pmatrix},$$

   which is a positive semi-definite matrix. Since it is a diagonal matrix, its eigenvalues are directly given by the diagonal elements. Due to the singularity of the Hessian, $\bar{\mathbf{w}}$ is said to be a degenerated critical point and a clear conclusion about its nature cannot be made. However, we can use the Taylor expansion of H_S to analyze its behavior in a neighborhood of $\bar{\mathbf{w}}$. Consider increments $\mathbf{h} = (h_1, h_2, h_3)^{\mathrm{T}}$ where $h_i, \ i = 1, \ldots, 3$, is small. We can write

$$H_S(\bar{\mathbf{w}} + \mathbf{h}) = H_S(\bar{\mathbf{w}}) + \mathbf{h}^{\mathrm{T}} \nabla H_S(\bar{\mathbf{w}}) + \mathbf{h}^{\mathrm{T}} \nabla^2 H_S(\bar{\mathbf{w}}) \mathbf{h} + o(\|\mathbf{h}\|^2). \quad (2.10)$$

   For very small $\|\mathbf{h}\|$, one may neglect $o(\|\mathbf{h}\|^2)$ and write

$$H_S(\bar{\mathbf{w}} + \mathbf{h}) - H_S(\bar{\mathbf{w}}) \approx \mathbf{h}^{\mathrm{T}} \nabla^2 H_S(\bar{\mathbf{w}}) \mathbf{h}. \quad (2.11)$$

   Now, if the Hessian were positive definite (all positive eigenvalues), then for any $\mathbf{h}$, the quadratic form $\mathbf{h}^{\mathrm{T}} \nabla^2 H_S(\bar{\mathbf{w}}) \mathbf{h}$ would be positive and $\bar{\mathbf{w}}$ would be a strict local minimum. However, it is easy to see that there are increments $\mathbf{h}$ such that $\mathbf{h}^{\mathrm{T}} \nabla^2 H_S(\bar{\mathbf{w}}) \mathbf{h} = 0$; these are of the form $\mathbf{h} = (h_1, 0, 0)$. But in this case, $\bar{\mathbf{w}} + \mathbf{h}$ belongs to the positive w_1 axis where H_S is constant. Along any other $\mathbf{h}$ directions, the quadratic form is positive. This means that $\bar{\mathbf{w}}$, or more precisely, the whole positive w_1 axis, is in fact an entropy minimum.

2. $\mu_{11} = 0.5$ and $w_1 > 0$
   The Hessian now becomes

$$\nabla^2 H_S(\bar{\mathbf{w}}) \approx \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{0.2641}{w_1^2} & 0 \\ 0 & 0 & \frac{-0.1377}{w_1^2} \end{pmatrix}. \tag{2.12}$$

   This matrix is indefinite, because it has positive and negative eigenvalues. This means that there are directions such that $\bar{\mathbf{w}}$ is a minimum and directions such that $\bar{\mathbf{w}}$ is a maximum (and of course, as discussed above, directions where H_S remains constant). These critical points are saddle points.

This analysis shows that the discrete MEE principle applied to hyperplane learning, is even less general than in the unidimensional case. In fact, while for Stoller split problems the minimum of entropy changes to maximum as the classes get closer [17], in the bivariate case the minimum may change to a saddle point, which brings about further difficulties when applying an optimization strategy.

2.2.3 Minimum Distance for Gaussian Classes

It is worth asking when are the Gaussian classes no longer "distant" and the minimum of H_S turns into a maximum. This can be studied using the eigenvalues of the Hessian matrix. In fact, $\nabla^2 H_S(\bar{\mathbf{w}})$ is always a diagonal matrix with one zero entry, an always positive entry, and a third entry that changes sign as the classes get closer (as previously illustrated); these entries are the eigenvalues of the matrix. We can, therefore, determine the minimum distance yielding a minimum of H_S at $(w_1, 0, 0)^{\mathrm{T}}$ for $w_1 \neq 0$, by inspecting when the sign-changing eigenvalue changes of sign. With $\mu_{11} = -\mu_{\text{-}11}$, $\mu_{12} = \mu_{\text{-}12} = 0$ and $\Sigma_1 = \Sigma_2 = \sigma^2 I$, this eigenvalue can be written as a function of $d = \mu_{11}/\sigma$, which can be seen as a normalized half distance between the classes. The obtained expression is rather long and but it can be verified that the eigenvalue is positive if the following expression is positive:

$$\sqrt{2\pi} d (1 - F(d)) \ln\left(\frac{2F(d)}{1 - F(d)}\right) - e^{-\frac{d^2}{2}}, \tag{2.13}$$

where F is the cumulative distribution of $G_1(0, 1)$. The turning value is approximately $d = 0.7026$, which corresponds to a normalized distance between the classes of approximately 1.4052. This is precisely the same value encountered for the Stoller split problem [17].

2.2.4 Equal Error Probabilities as a Necessary Condition

We now prove that equal class error probabilities is a necessary condition to ensure that the optimal solution $\mathbf{w}^*$ is a critical point of error entropy. This is a multivariate version of Theorem 3 in [17].

Theorem 1. *In the two-class multivariate problem, if the optimal set of parameters* $\mathbf{w}^* = (w_1^*, \ldots, w_d^*, w_0^*)^T$ *of a separating line constitute a critical point of the error entropy then the error probabilities of each class at* $\mathbf{w}^*$ *are equal.*

Proof.
We start by noticing that the linear discriminant can be viewed has a one-dimensional classification problem. In fact, $\bar{z} = \mathbf{w}^{\mathrm{T}}\mathbf{x}$ is a projection of $\mathbf{x}$ onto $\mathbf{w}$. From an initial distribution represented by a density $g(\mathbf{x}) = qg_{X|\text{-}1}(\mathbf{x}) + pg_{X|1}(\mathbf{x})$ we get, on the projected space, the distribution of the projected data given by $f(\bar{z}) = qf_{\bar{z}|\text{-}1}(\bar{z}) + pf_{\bar{z}|1}(\bar{z})$. The parameter w_0 then works as a Stoller split: a given pattern is classified in $\mathcal{C}_1$ if $\bar{z} \geq w_0$. Thus, and from the results in [17], we can assert that $qf_{\bar{z}|\text{-}1}(\bar{z}) = pf_{\bar{z}|1}(\bar{z})$ at $\mathbf{w}^*$.
We rewrite the error probabilities of each class as

$$P_{\text{-}1} = q(1 - F_{\bar{z}|\text{-}1}(-w_0)), \tag{2.14}$$

$$P_1 = pF_{\bar{z}|1}(-w_0) \tag{2.15}$$

where $\bar{z} = \mathbf{w}^{\mathrm{T}}\mathbf{x}$. Thus

$$\frac{\partial P_{\text{-}1}}{\partial w_0} = -qf_{\bar{z}|\text{-}1}(-w_0) \tag{2.16}$$

$$\frac{\partial P_1}{\partial w_0} = pf_{\bar{z}|1}(-w_0) \tag{2.17}$$

From (1.4)

$$\frac{\partial H_S}{\partial P_t} = \ln\left(\frac{1 - P_{\text{-}1} - P_1}{P_t}\right) \qquad t \in \{\text{-}1, 1\}$$

From the chain rule and using the fact that $qf_{\text{-}1} = pf_1$ at $\mathbf{w}^*$ we see that

$$\frac{\partial H_S}{\partial w_0}(\mathbf{w}^*) = 0 \Leftrightarrow \tag{2.18}$$

$$\Leftrightarrow pf_{\bar{z}|1}(w_0^*)\left(\ln\left(\frac{1 - P_{\text{-}1} - P_1}{P_{\text{-}1}}\right) - \ln\left(\frac{1 - P_{\text{-}1} - P_1}{P_1}\right)\right) = 0 \tag{2.19}$$

$$\Leftrightarrow f_{\bar{z}|1}(w_0^*) = 0 \vee P_{\text{-}1} = P_1 \tag{2.20}$$

Note that $f_{\tilde{z}|1}(w_0^*) = 0$ if and only if the classes have distributions with disjoint supports (they are separable). But in this case $P_{\text{-}1} = P_1 = 0$. Thus, in both cases $P_{\text{-}1} = P_1$ is a necessary condition.

q.e.d.

If it were possible to compute the partial derivatives with respect to w_1 and w_2 one could also show whether or not the equal-error-probability condition is also sufficient. Two examples are given in Appendix .3 illustrating this Theorem.

Chapter 3

Perceptron with Continuous Errors

The continuous error distribution is characterized by formula (1.1). We consider two types of a.f.: linear and squashing functions. We also use two definitions of entropy, viz. Shannon entropy, $H_S(E)$, and Rényi's entropy, $H_{R_\alpha}(E)$, defined as in the Introduction. Our goal is again to study the *classifier problem* in light of the MEE principle. More precisely, we will study the behavior of the error entropy as we vary the parameters of $\varphi_{\mathbf{w}}(x)$ (we often simply denote by $\varphi(x)$) and investigate if the theoretical optimal solution corresponds to an error distribution with minimum entropy.

3.1 The Split-Type Setting

We start by considering perceptrons with only one adjustable parameter; that is, the perceptron is trained to find a split point in the error distribution, as in the Stoller split setting. The only difference is that now we have a continuous error distribution. In the following we consider the cases of linear and squashing activation functions.

3.1.1 Linear Activation Function

Consider two uniform overlapping classes defined by the densities[1]

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

with $a < c < b < d$ and $\varphi(x) = x - w$, where w is a threshold. Note that, if $\varphi(x) \geq 0 \iff x \geq w$ we classify x as $\mathcal{C}_1$, otherwise we classify as $\mathcal{C}_{\text{-}1}$. One easily derives

$$f_{Y|\text{-}1}(-1-e) = \frac{1}{b-a} I_{[w-b-1,w-a-1]}(e) \tag{3.2}$$

$$f_{Y|1}(1-e) = \frac{1}{d-c} I_{[w-d+1,w-c+1]}(e) \tag{3.3}$$

The final configuration of the transformed (shifted) distributions is dependent on the values of $c-a$ and $d-b$ (in some cases the optimal solution would need two splits). Also, E is not necessarily constrained to the interval $[-2, 2]$. We analyze the case where the final distributions are such that $w-d+1 < w-b-1 < w-c+1 < w-a-1$, that is we assume an overlap in the interval $[w-b-1,\ w-c+1]$. In this case,

$$\begin{aligned} H_S(E) = -\Bigg[& \int_{w-d+1}^{w-b-1} \frac{p}{d-c} \ln\left(\frac{p}{d-c}\right) de + \\ & \int_{w-b-1}^{w-c+1} \left(\frac{p}{d-c} + \frac{q}{b-a}\right) \ln\left(\frac{p}{d-c} + \frac{q}{b-a}\right) de + \\ & \int_{w-c+1}^{w-a-1} \frac{q}{b-a} \ln\left(\frac{q}{b-a}\right) de \Bigg] \end{aligned} \tag{3.4}$$

Thus:

$$\begin{aligned} H_S(E) = & \frac{p(d-b-2)}{d-c} \ln\left(\frac{p}{d-c}\right) + \\ & \left(\frac{p}{d-c} + \frac{q}{b-a}\right) \ln\left(\frac{p}{d-c} + \frac{q}{b-a}\right) (b-c+2) + \\ & \frac{q(c-a-2)}{b-a} \ln\left(\frac{q}{b-a}\right) \end{aligned} \tag{3.5}$$

which does not dependent on w! Hence, the MEE principle doesn't work with linear a.f., which anyway isn't the most appropriate for classification problems.

[1] $I_{[a,b]}(x)$ is the indicator function with value 1 whenever $x \in [a, b]$ and 0 otherwise.

3.1.2 Squashing Activation Function

We take as squashing function of a single perceptron the popular and mathematically easy to manipulate tanh function. In order to derive the error pdf, we start by recalling that H_S can be decomposed as a sum of error sub-entropies as stated in formula (1.3). For the Rényi's entropy one derives

$$H_{R_\alpha} = \frac{1}{1-\alpha} \ln \int_{-\infty}^{\infty} [f_E(e)]^\alpha \, de \tag{3.6}$$

$$= \frac{1}{1-\alpha} \log \ln \left[\int_{-2}^{0} [q f_{\text{-}1}(e)]^\alpha \, de + \int_{0}^{2} [p f_1(e)]^\alpha \, de \right] \tag{3.7}$$

Although Rényi's entropy is not decomposable as a sum of class sub-entropies, the minimization problem can be transformed into an equivalent problem where a sum of two positive quantities (each exclusively related to each class) appears. As an example, for the special case with $\alpha = 2$, the minimization of H_{R_2} is equivalent to the maximization of

$$V_{R_2} = \exp(-H_{R_2}) = \int_{-2}^{0} [q f_{\text{-}1}(e)]^2 \, de + \int_{0}^{2} [p f_1(e)]^2 \, de \tag{3.8}$$

This decomposition is an important property of MEE for classification and emphasizes the difference between classification and regression (as previously discussed). The same decomposition appears for multi-class problems. In fact, whenever a pdf $f(x)$ can be written as

$$f(x) = \sum_i a_i f_i(x)$$

with $\sum_i a_i = 1$ and the supports D_i of the pdf's $f_i(x)$ are such that $D_i \cap D_j = \varnothing$, $\forall i \neq j$, then the Shannon's entropy H_f associated with f is given by

$$H_f = -\sum_i a_i \ln(a_i) + \sum_i a_i H_{f_i},$$

where H_{f_i} is the Shannon's entropy associated to f_i. This applies to multi-class problems whenever an 1-out-of-C coding is used. An equivalent decomposition appears for V_{R_2}.

For two-class problems with squashing a.f. we also need to use the well-known theorem of r.v. transformation:

Theorem 2. *Let $f(x)$ be the pdf of the r.v. X. Assume $\varphi(x)$ to be monotonic and differentiable and suppose $\varphi'(x) \neq 0 \ \forall x$. If $g(y)$ is the density of $Y = \varphi(X)$ then*

$$g(y) = \begin{cases} \dfrac{f(\varphi^{-1}(y))}{|\varphi'(\varphi^{-1}(y))|}, & \inf \varphi(x) < y < \sup \varphi(x) \\ 0, & otherwise \end{cases} \tag{3.9}$$

where $x = \varphi^{-1}(y)$ is the inverse function of $y = \varphi(x)$.

Note that our aim is to compute the density of E, which is a transformation of the input X. For the split-type case, $\varphi(x) = \tanh(x-w)$, is a differentiable and strictly increasing transformation, where w acts as the split point. We thus have

$$\varphi'(x) = 1 - \tanh^2(x - w) \neq 0 \ \forall x \tag{3.10}$$

$$\varphi^{-1}(y) = w + \operatorname{arctanh}(y) \tag{3.11}$$

$$\varphi'(\varphi^{-1}(y)) = 1 - y^2 \tag{3.12}$$

We now study the special cases of uniform and Gaussian distributed input data.

Uniform Classes

Suppose that the two classes have inputs described by two overlapping uniform densities as in (3.1). Making use of Theorem 2 one derives

$$f_{Y|\text{-}11}(-1-e) = \frac{-1}{(b-a)e(2+e)} I_{[-1-\tanh(b-w),-1-\tanh(a-w)]}(e) \tag{3.13}$$

$$f_{Y|1}(1-e) = \frac{1}{(d-c)e(2-e)} I_{[1-\tanh(d-w),1-\tanh(c-w)]}(e) \tag{3.14}$$

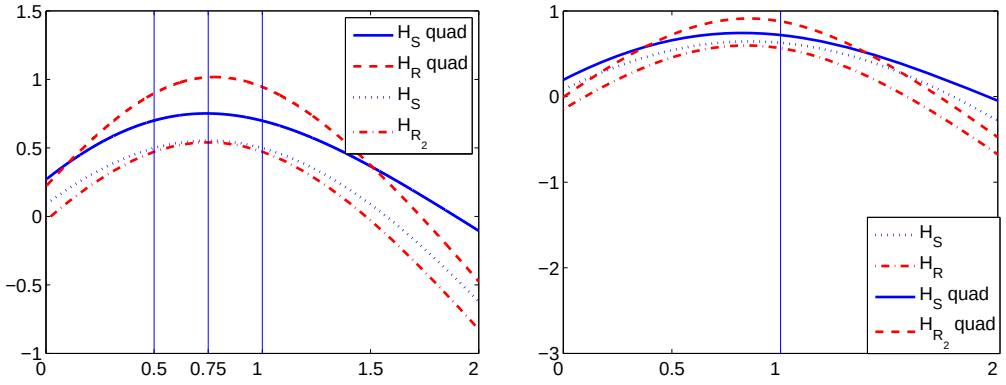
Thus, from (1.3), we obtain

$$\begin{aligned} H_S = q \Bigg[& \frac{2\ln\left(\frac{|e|}{2+e}\right)\ln\left(\frac{-1}{(b-a)e(2+e)}\right) + 4\,\mathrm{dilog}\left(\frac{2+e}{2}\right)}{4(b-a)} + \\ & + \frac{\ln|e|\ln\left(\frac{|e|(2+e)^2}{16}\right) + 2\ln(2)^2 - \ln(2+e)^2}{4(b-a)} \Bigg]_{-1-\tanh(b-w)}^{-1-\tanh(a-w)} + \\ & + p \Bigg[\frac{2\ln\left(\frac{e}{|e-2|}\right)\ln\left((d-c)e(2-e)\right) + 4\,\mathrm{dilog}\left(\frac{e}{2}\right)}{4(d-c)} + \\ & + \frac{\ln|e-2|\ln\left(\frac{e^2|e-2|}{16}\right) + 2\ln(2)^2 - \ln(e)^2}{4(d-c)} \Bigg]_{1-\tanh(d-w)}^{1-\tanh(c-w)} + H_S(T) \end{aligned} \tag{3.15}$$

and for Rényi's entropy

$$\begin{aligned} V_{R_2} = -\frac{q^2}{4} & \left[\frac{2 + e(2+e)\ln\left(\frac{|e|}{2+e}\right) + 2e}{(b-a)^2(2+e)e} \right]_{-1-\tanh(b-w)}^{-1-\tanh(a-w)} + \\ & + \frac{p^2}{4} \left[\frac{2 + \ln\left(\frac{e}{|e-2|}\right) e(e-2) - 2e}{(d-c)^2(e-2)e} \right]_{1-\tanh(d-w)}^{1-\tanh(c-w)} \end{aligned} \tag{3.16}$$

Figure 3.1 shows H_S and H_{R_2} as a function of w using exact and approximate computation of the integrals. The deviation of the approximate solution from the exact one is attributed to the fact that inaccuracies are unavoidable



(a) $c = 0.5,\ d = 1.5$ (b) $c = 0.5,\ d = 2$

Figure 3.1: Shannon and Rényi entropies as a function of w. Dashed lines are obtained with exact computations, while solid lines use approximated computation (quadrature) of the integrals.

in the numerical evaluation of the integrals near the diverging tails of the integrands. Class $\mathcal{C}_{-1}$ is fixed to $[a, b] = [0, 1]$ and $p = q = 1/2$. In Figure 3.1(a), where the classes have equal support width, the optimal split is any point in the interval $[0.5, 1]$. Both Shannon and Rényi's entropies have a maximum at $w = 0.75$. This is in direct contradiction with the MEE criterion, which states that w should be chosen so as to minimize the error entropy. The particular choice of this split can be explained by the fact that this is the split point providing equal class error probability. This was already encountered in the discrete entropy case [17]. In general, one can prove the following. Let $a = 0 \leq c \leq b \leq d = c + k$, where $k \in \mathbb{R}$ controls the support width of $\mathcal{C}_1$. For $k \geq b - a$ the optimal split point for the $b = 1$ setting occurs obviously at $w = 1$, since it will correspond to $\min Pe$. For Shannon entropy

$$\frac{dH_S}{dw} = \frac{k \ln\left(\frac{\cosh^2(w)}{\cosh^2(b-w)}\right) + b \ln\left(\frac{\cosh(c-w)^2}{\cosh(c+k-w)^2}\right)}{-2bk}, \tag{3.17}$$

$$\frac{d^2H_S}{dw^2} = -\frac{\frac{k\sinh(b)}{cosh(w)cosh(b-w)} + b\left(\frac{\sinh(c+k-w)}{\cosh(c+k-w)} - \frac{\sinh(c-w)}{\cosh(c-w)}\right)}{bk}. \tag{3.18}$$

If we take $k = b$, and thus, both classes have equal support width, we get

$$\frac{dH_S}{dw}\left(\frac{b+c}{2}\right) = 0 \quad \wedge \quad \frac{d^2H_S}{dw^2}\left(\frac{b+c}{2}\right) < 0$$

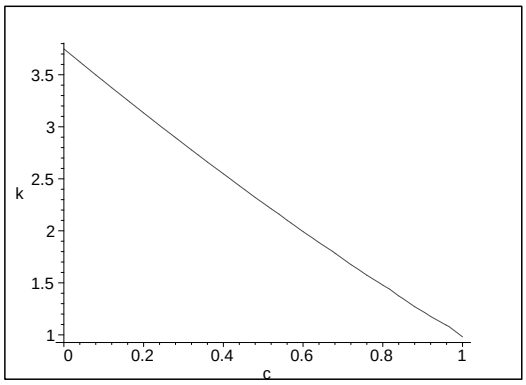


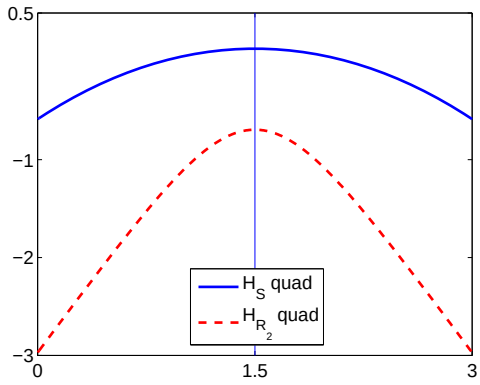
Figure 3.2: Contour level (zero) of $\frac{dH_S}{dw}$ as a function of c and k.

which means that $(b+c)/2$ is a maximizer of H_S.
A rather unexpected behavior appears when the support of class $\mathcal{C}_1$ is increased. In Figure 3.1(b), where $[c,d] = [0.5, 2]$, the optimal split moves toward an unique point, $w = 1$. However, both Shannon and Rényi's entropies fail to identify it (now, the maximum is at $w \approx 0.859$ and $w \approx 0.841$, respectively). Note that in this case, the class error probabilities are not equal. We know that in the discrete case a necessary condition for the optimal split corresponding to the entropy extrema is that the class error probabilities are equal [17]. However, in the present case this correspondence is not valid. This can be seen by first answering the question: is there any combination of c and k which yields $w = 1$ as the optimal solution? Figure 3.2 answers this question by showing the solution of $\frac{dH_S}{dw} = 0$ for $w = 1$ and $b = 1$. This figure tells us that k has to be greater than 1 and furthermore as c decreases, a higher k is needed. As an example, we may see that while the setting $[a,b] = [0,1]$ and $[c,d] = [1,2]$ has a maximum at $w = 1$, the setting $[a,b] = [0,1]$ and $[c,d] = [1,1.9]$ does not. This also contradicts the "equal-error necessary condition" hypothesis, because for $c < 1 < k$ the error probabilities are *not* equal (for example, for $c = 0.8$ one should have $k \approx 1.48$). Can these behaviors be attributed to the fact that the uniform pdf is not continuous? We proceed to analyzing the case of two Gaussian classes, where we will find the same behavior.

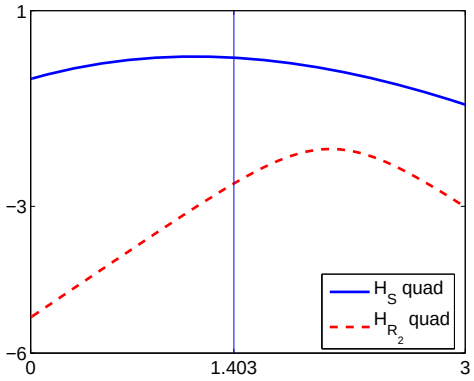
Gaussian Classes

Let us now consider the case where the input distributions are Gaussian

$$f_{X|\text{-}1}(x) \sim N(\mu_{\text{-}1}, \sigma^2_{\text{-}1}) \qquad f_{X|1}(x) \sim N(\mu_1, \sigma^2_1)$$



(a) $\mu_1 = 3,\ \sigma_1 = 1$



(b) $\mu_1 = 3,\ \sigma_1 = 2$

Figure 3.3: Shannon and Rényi's entropies as a function of w.

Applying Theorem 2 one easily gets

$$f_{Y|t}(t-e) = \frac{\exp\left(-\frac{1}{2}\left(\frac{\operatorname{arctanh}(t-e)-(\mu_t-w)}{\sigma_t}\right)^2\right)}{\sqrt{2\pi}\sigma_t\ e(2t-e)} I_{[t-1,t+1]}(e) \tag{3.19}$$

Figure 3.3 shows H_S and H_{R_2} as a function of w using approximate computation of the integrals (there is no closed form for the integrals). Class $\mathcal{C}_{\text{-}1}$ is fixed to $(\mu_{\text{-}1}, \sigma_{\text{-}1}) = (0, 1)$ and $p = q = 1/2$. In Figure 3.3(a), where the class means just differ in location, the optimal split is the middle point between the class means, $w^* = 1.5$. Both entropies find this point as a maximum. Moreover, in Figure 3.3(b), where $(\mu_1, \sigma_1) = (3, 2)$, the optimal split changes to $w \approx 1.403$. Both entropies fail to identify this point. Again the error probabilities for each class are equal in the former case, while this

does not happen in the latter case.

These theoretical behaviors of both Shannon and Rényi's entropies (maximum instead of minimum and displaced from the $\min_W Pe_\Phi$ position) raise the natural question: how is it possible that the MEE principle works well in practice? There are two main aspects that differentiate the preceding theoretical analysis from the practical implementation. The first one is related to the learning machine's flexibility/complexity. In fact, in the preceding examples the perceptron was allowed only a sliding split that basically sets the location of the a.f.. Whether a more flexible activation performs better is investigated in the following section. Secondly, as the true class distributions are not known, the error distribution cannot be computed using tools like Theorem 2. A kernel density estimator is used in practice and its effect is also investigated in a forthcoming section.

3.2 The Perceptron Setting

We now assume $\varphi(x) = \tanh(w_1 x - w_0)$. That is, instead of a split-type setting, controlled by w_0 imposing a simple sliding of $\varphi(x)$, we now have a more realistic perceptron setting with a parameter, w_1, controlling the *function shape* of $\varphi(x)$ (in fact, the steepness of φ). In particular, $\varphi(x) \to H(x)$ as $w_1 \to +\infty$, where $H(x)$ is the threshold a.f.. We also assume that $w_1 > 0$ since for $w_1 = 0$, no adaptation would be possible and if $w_1 < 0$, φ would perform a swapped classification. Using Theorem 2 one derives

$$\varphi'(x) = w_1(1 - \tanh^2(w_1 x - w_0)) \neq 0 \ \forall x \tag{3.20}$$

$$\varphi^{-1}(y) = \frac{1}{w_1}(w_0 + \operatorname{arctanh}(y)) \tag{3.21}$$

$$\varphi'(\varphi^{-1}(y)) = w_1(1 - y^2) \tag{3.22}$$

We repeat the previous analysis for uniform and Gaussian classes.

Uniform Classes

The error pdf's for uniform classes are again obtained using Theorem 2.

They look very similar to the previous ones

$$f_{Y|\text{-}1}(-1-e) = \frac{-1}{w_1(b-a)e(2+e)} I_{[-1-\tanh(w_1 b - w_0), -1-\tanh(w_1 a - w_0)]} \quad (3.23)$$

$$f_{Y|1}(1-e) = \frac{1}{w_1(d-c)e(2-e)} I_{[1-\tanh(w_1 d - w_0), 1-\tanh(w_1 c - w_0)]} \quad (3.24)$$

Entropy is now a function of two variables, w_1 and w_0. Figure 3.4 shows the surface of H_S and its contours. Two examples are shown with $[a, b] = [0, 1]$ and $[c, d] = [0.2, 1.2]$ in Figure 3.4(a) and $[c, d] = [0.9, 1.9]$ in Figure 3.4(b). The grids for w_1 and w_0 are chosen so that they are able to display the optimal solutions, namely the middle points of the overlapping intervals. We see that both surfaces have a maximum. Analyzing the more informative contour plots, we encounter interesting behaviors. Let us first analyze the case where the overlapping region is $[0.2, 1]$ (Figure 3.4(a)). Any split in this interval is optimal. In particular, the "middle" optimal split (the one corresponding to equal class error probabilities) corresponds to the $w_0/w_1 = 0.6$ line. This line (solid line) is represented over the contour plot together with the $w_0/w_1 = 0.2$ and $w_0/w_1 = 1$ dashed lines, also achieving $\min Pe$. The solid line appears to pass at the location of the maximum as can be more clearly seen in the zoomed image (elliptical axes). However, instead of yielding the whole $w_0/w_1 = 0.6$ line as a solution, i.e., instead of exhibiting a straight "ridge", the entropy surface exhibits a single peak. In the bottom figures we encounter a similar behavior. It is interesting to see that in this case a lower value for w_1 is obtained. In fact one observes a dependency between the steepness of the a.f. and the amount of overlap, with an increased overlap requiring an increased steepness of the a.f.. Consider $H_S = H_S(w_1, w_0)$ and $a = 0 \leq c \leq b \leq d = c + b$ (classes with equal-length support). Then[2]

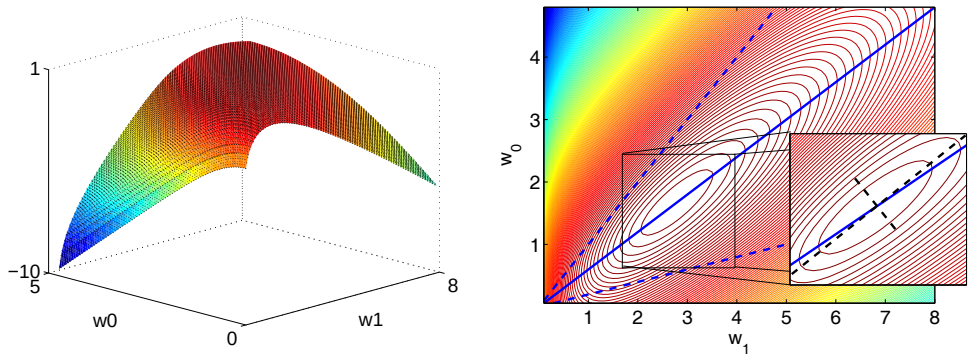
$$\frac{\partial H_S}{\partial w_0}\left(w_1, w_1 \frac{b+c}{2}\right) = 0 \quad (3.25)$$

This means that the middle point of the overlapped region is a candidate for an extremum. Its nature can be studied using the second order information given by the Hessian. We then verify that

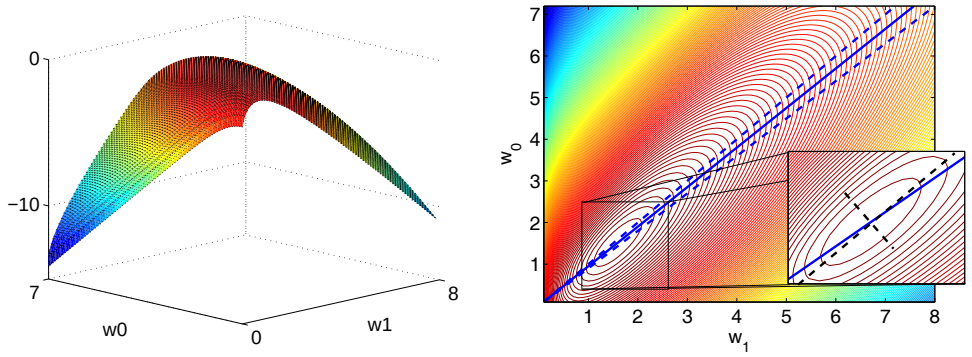
$$\frac{\partial^2 H_S}{\partial w_0^2}\left(w_1, w_1 \frac{b+c}{2}\right) < 0 \quad (3.26)$$

$$\frac{\partial^2 H_S}{\partial w_1 \partial w_0}\left(w_1, w_1 \frac{b+c}{2}\right) = \frac{\partial^2 H_S}{\partial w_0 \partial w_1}\left(w_1, w_1 \frac{b+c}{2}\right) > 0 \quad (3.27)$$

[2]The partial derivative with respect to w_1 is intractable.



(a) $c = 0.2,\ d = 1.2$



(b) $c = 0.9,\ d = 1.9$

Figure 3.4: Surfaces and contour plots of $H_S(w_1, w_0)$ for different values of $[c, d]$.

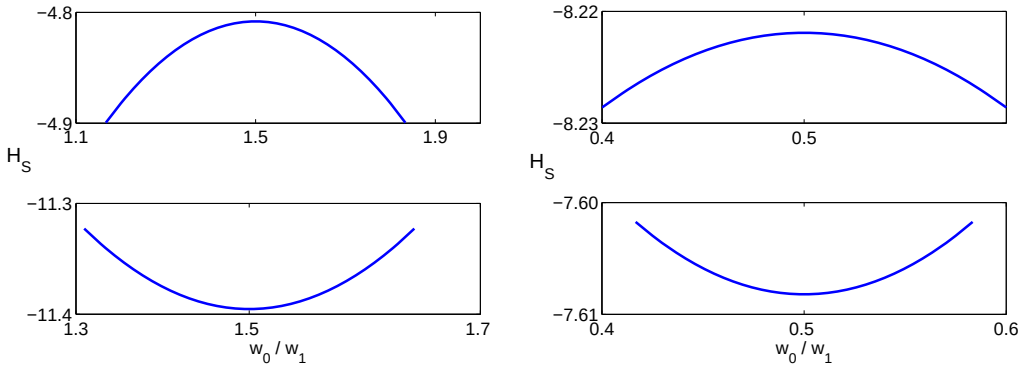
The expression for $\partial^2 H_S / \partial w_1^2$ is intractable. With this information one can be sure that if the critical point $(w_1, w_1(b + c)/2)$ is not a saddle point, then it is a maximum.

Gaussian Classes

The Gaussian transformed pdf's are derived as

$$f_{Y|t}(t-e) = \frac{\exp\left(-\frac{1}{2}\left(\frac{\operatorname{arctanh}(t-e)-(w_1\mu_t-w_0)}{w_1\sigma_t}\right)^2\right)}{\sqrt{2\pi}w_1\sigma_t\ e(2t-e)} I_{[t-1,t+1]}(e) \tag{3.28}$$

In Figure 3.5, we specified a value for w_1 (the steepness parameter) and let w_0 vary in a way such that the optimal solution is displayed. Thus, H_S is plotted as a function of w_0/w_1. Without loss of generality, we set the "left class" with $(\mu_{-1}, \sigma_{-1}) = (0, 1)$. Figure 3.5(a) refer to the setting

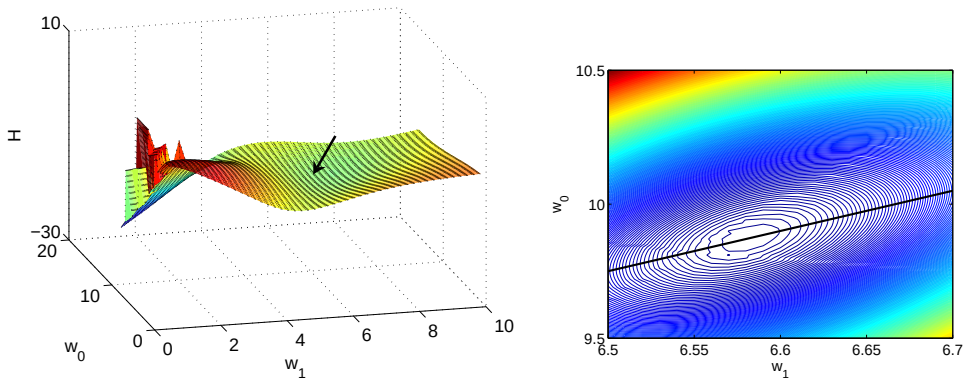


(a) $w_1 = 3$ at top; $w_1 = 7$ at bottom (b) $w_1 = 10$ at top; $w_1 = 15$ at bottom

Figure 3.5: $H_S(w_1, w_0)$ for fixed values of w_1 and different locations of the Gaussians.

$(\mu_1, \sigma_1) = (3, 1)$ with corresponding optimal split at $x^* = 1.5$. We observe that the increase of w_1 causes H_S to change from a maximum to a minimum at x^*. In Figure 3.5(b), where $(\mu_1, \sigma_1) = (1, 1)$ and $x^* = 0.5$, we observe that the increase of overlap between the classes requires an higher value of w_1 to perform the same change.

These results suggest the need of using function shaping parameters, as is the case with multilayer perceptrons, in order to get an entropy minimum. Figure 3.6 shows H_S as a function of (w_1, w_0), where one can identify the previous behaviors. Note that the minima attained for small values of w_1 and high values of w_0 do not correspond to an optimal solution (due to the relation $w_1 x = w_0$). Nonetheless, a local minimum is attained at $(w_1, w_0) \in [6.5,\ 6.7] \times [9.5,\ 10.5]$ as shown in the contour plot. As before, the solid line



(a) (b)

Figure 3.6: Surface and contour plot of $H_S(w_1, w_0)$.

represents the set of solutions $w_0/w_1 = 1.5$. Again, the line appears to pass through the minimum. Unfortunately, due to the complexity of the formulas, a functional analysis similar to the one performed for uniform classes, is not possible. We were also able to observe that if the classes get closer, the same behavior is obtained, namely the need of a higher w_1 in order to obtain the minimum.

3.3 Estimating the Error Density

There is an essential difference between the theoretical MEE and how it is implemented. In fact, the input distributions are usually unknown which makes it impossible to use Theorem 2 to determine the exact error distributions. A method to estimate the error pdf's is then used. The usual method is the kernel density estimator (kde) also known as Parzen window estimator. Given a dataset $x_1, x_2, \ldots, x_N$ an estimate $\hat{f}(x)$ of the pdf $f(x)$ is given by

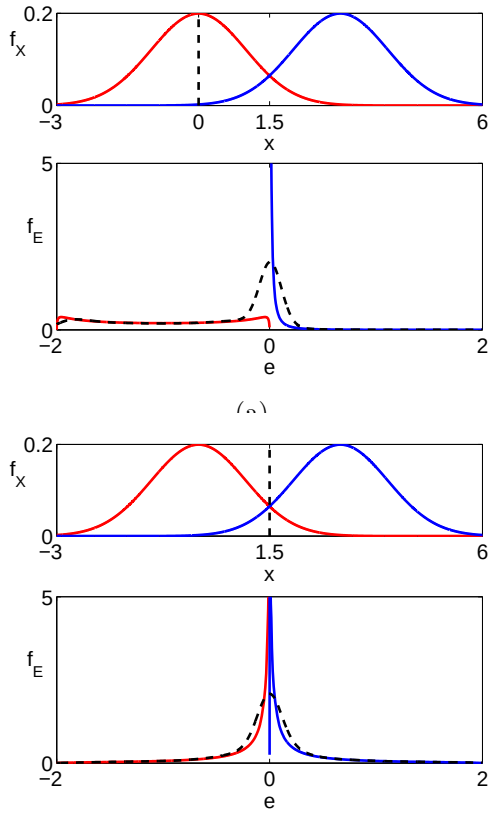
$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \tag{3.29}$$

where K is a kernel function and h is the kernel bandwidth (smoothing parameter). The standardized Gaussian pdf enjoys desirable properties as kernel function, thereby is popularly used and is the one that we consider. To analyze the kde impact on the MEE principle for classification we performed the following experiment. Two Gaussian classes were generated with 10000

data points each. Class $\mathcal{C}_{\text{-}1}$ was always centered at the origin and its standard deviation was 1. Class $\mathcal{C}_1$ was generated in two different settings differing from $\mathcal{C}_{\text{-}1}$ only in its location: $\mu_1 \in \{1, 3\}$. We then applied the following transformation

$$e_i = t - \tanh(x_i - w), \qquad x_i \in \mathcal{C}_t,\ t \in \{-1, 1\} \tag{3.30}$$

for a grid of w values. Figure 3.7 shows the theoretical and an instance of the practical error pdf's. Note the smoothing imposed by the kde. Entropy



(b)

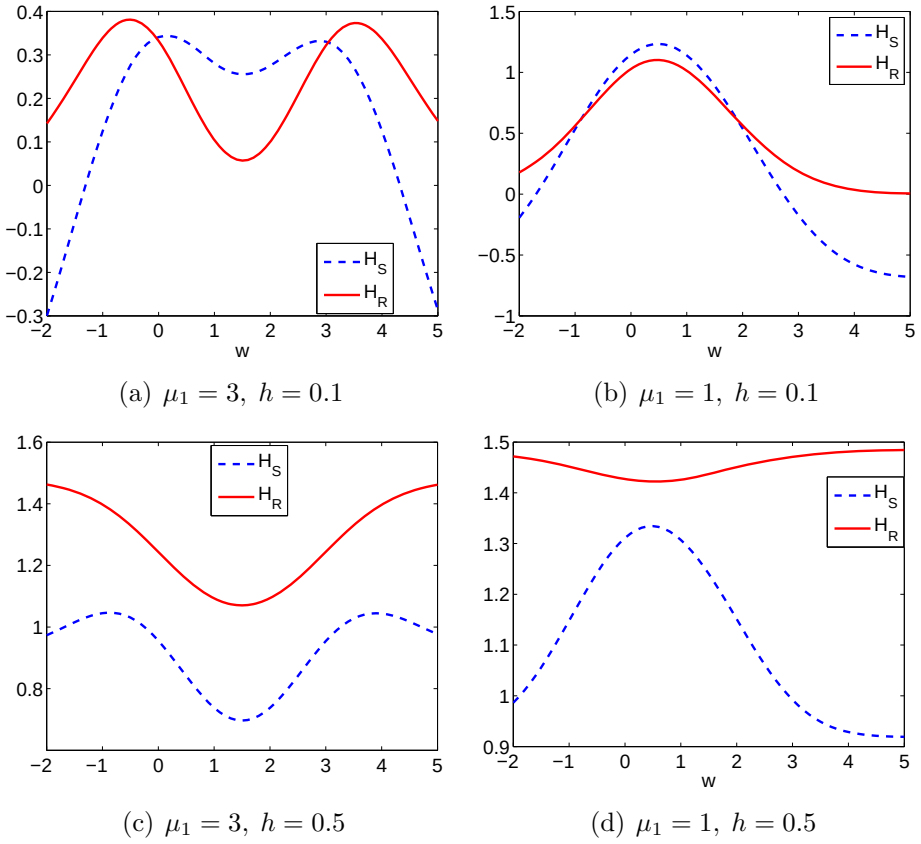
Figure 3.7: The kde smoothing effect. The top figures show the class pdf's with the split location (dashed vertical line). The bottom figures show the theoretical (solid line) and kde (dashed line) error pdf's for the corresponding split.

was finally estimated using the following relations [18, 5]

$$H_S = \mathbb{E}\{\ln f(e)\} \approx \frac{1}{N}\sum_{i=1}^{N} \ln f(e_i) \approx \frac{1}{N}\sum_{i=1}^{N} \ln \hat{f}(e_i) \tag{3.31}$$

$$H_{R_2} = -\ln \frac{1}{\sqrt{2}hN^2} \sum_{i=1}^{N}\sum_{j=1}^{N} K\left(\frac{e_i - e_j}{\sqrt{2}h}\right) \tag{3.32}$$

Note that, by using enough data and a proper h, both approximations are reliable [19]. The joint effect of varying the smoothing parameter h and of increasing/decreasing the overlap between the classes is shown in Figure 3.8 for both Shannon and Rényi's entropies as functions of the split point w. From left to right we increase the overlap while from top to bottom we increase h. We see that the increase of h implies a change from a maximum



(a) $\mu_1 = 3,\ h = 0.1$ (b) $\mu_1 = 1,\ h = 0.1$

(c) $\mu_1 = 3,\ h = 0.5$ (d) $\mu_1 = 1,\ h = 0.5$

Figure 3.8: Effect of the kde in Shannon's and Rényi's error entropies.

to a minimum. Also, this optimal extreme gradually becomes "less local". The increase of overlap mainly requires greater values of h to obtain the same behavior. It is now clear the impact of the kde when using the MEE principle, even for this "worst" case, the split-type setting, where we have previously shown the sole existence of a maximum. Note that we cannot say that any one of the used entropies is better in some sense than the other, because they clearly work with different values for h (higher for Shannon entropy). In fact, they are quite similar if we compare Rényi's for $h = 0.5$ with Shannon's for $h = 1.0$ when $\mu_1 = 3$. Let us examine the expression of the empirical entropy. We use Rényi's expression for simplicity and to better emphasize the relations. The minimization of (3.32) is equivalent to the maximization of

$$\hat{V}_{R_2} = \frac{1}{\sqrt{2}hN^2} \sum_{i=1}^{N} \sum_{j=1}^{N} K\left(\frac{e_i - e_j}{\sqrt{2}h}\right) \tag{3.33}$$

Let $s = \frac{e_i - e_j}{\sqrt{2}h}$, $c = \frac{1}{\sqrt{2}hN^2}$ and $c_t = \frac{1}{\sqrt{2}hN_t^2}$ for $t \in \{-1, 1\}$ where N_t is the number of samples from class $\mathcal{C}_t$. Then, if K is symmetrical about the origin (as is the Gaussian kernel) we may write

$$\hat{V}_{R_2} = c \sum_{i \in \mathcal{C}_{\text{-}1}} \sum_{j \in \mathcal{C}_{\text{-}1}} K(s) + c \sum_{i \in \mathcal{C}_{\text{-}1}} \sum_{j \in \mathcal{C}_1} K(s) + c \sum_{i \in \mathcal{C}_1} \sum_{j \in \mathcal{C}_{\text{-}1}} K(s) + c \sum_{i \in \mathcal{C}_1} \sum_{j \in \mathcal{C}_1} K(s) \tag{3.34}$$

$$= \left(\frac{N_{\text{-}1}}{N}\right)^2 c_{\text{-}1} \sum_{i \in \mathcal{C}_{\text{-}1}} \sum_{j \in \mathcal{C}_{\text{-}1}} K(s) + \left(\frac{N_1}{N}\right)^2 c_1 \sum_{i \in \mathcal{C}_1} \sum_{j \in \mathcal{C}_1} K(s) + 2c \sum_{i \in \mathcal{C}_1} \sum_{j \in \mathcal{C}_{\text{-}1}} K(s) \tag{3.35}$$

$$= \hat{q}^2 \hat{V}_{R_2|\text{-}1} + \hat{p}^2 \hat{V}_{R_2|1} + 2c \sum_{i \in \mathcal{C}_1} \sum_{j \in \mathcal{C}_{\text{-}1}} K(s) \tag{3.36}$$

Entropy is, therefore, decomposed as a weighted sum of the error entropies for each class (as in the theoretic derivation) plus a term that relates the errors of one class with those of the other. For a small h this interference term is also small and $\hat{V}_{R_2}$ will be close to V_{R_2}. For large h the interference term will be large and the smoothing effect displayed in Figure ref will show up and gives rise to an entropy minimum. This behavior has been observed in the many experiments we have performed.

Chapter 4

Conclusions

The application of the MEE principle to error distributions in two-class problems was analyzed using a single perceptron. The main point of this analysis was to clarify how this information theoretic principle, that has the virtue of being based on the whole error pdf (and not a single measure of it), would cope with the classifier problem. The main motivation for this analysis was the good performance of MEE in many practical problems, often superior to the performance attained by using the ubiquitous MSE principle, in rigorously controlled experiments. This suggests that MEE is indeed a good principle for dealing with the classifier problem, i.e., for attaining the $\min_W Pe_\Phi$ allowed by the family of functions implemented by the classifier. In spite of the fact that we restricted the analysis to such a simple classifier as the perceptron, the theoretical analysis revealed nonetheless to be rather intricate in many circumstances. It did also reveal a large spectrum of behaviors that MEE can be expected to exhibit in practical problems. In what concerns the application of MEE to threshold-type machines (the discrete error case) applied to hyperplane learning, the analysis revealed that the MEE principle is even less general than in the unidimensional Stoller split case. In fact, while in Stoller split problems the minimum of entropy may change to a maximum for input data distributions that are close to each other, in the bivariate case and *a fortiori* in multivariate cases, the minimum may change to a saddle point, which brings about further difficulties when applying an optimization strategy. Therefore, for threshold-type machines the MEE principle should only be applied to well separated distributions. Some quantification of well-separateness for some univariate distributions can be found in [17]. The present work confirmed the same separating value for the bivariate case and Gaussian distributions. We were also able to prove

for the multivariate case that equal class error probabilities is a necessary condition for MEE to work. In what concerns the application of MEE to perceptrons having continuous activation functions (the continuous error case) after demonstrating that MEE cannot work with linear functions we investigated the use of a squashing function as activation function. We have shown that for MEE to work in this case, and using the true error pdf, there must exist parameters controlling the squashing function shape. A single location parameter (split-type setting) is not enough. This is of course not an important restriction at all, because in real practical problems there are many function shaping parameters. Moreover, and this is an important result, we have shown that by using Parzen window estimation of the pdf (given that in general practice the true error pdf is unknown) we are in fact using a smoothed version of the error pdf that clearly helps in setting the minimum error entropy at the optimal parameter vector corresponding to $\min_W Pe_\Phi$; as a matter of fact, Parzen window estimation often changes a *theoretical* error entropy maximum into a *practical* error entropy minimum.

.1 Example where MSE Fails

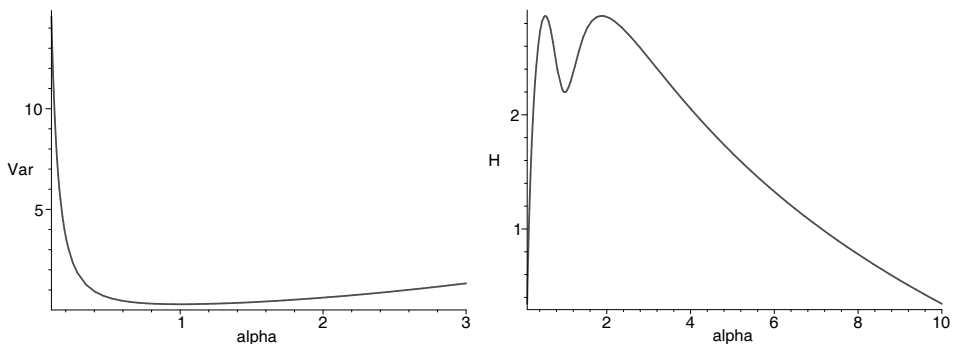
Let us suppose that E has a continuous distribution described by the following pdf

$$f(e) = \frac{1}{4}\left[Tr(e,0,\alpha) + Tr(e,-\alpha,0) + Tr(e,0,1/\alpha) + Tr(e,-1/\alpha,0)\right],$$

a sum of triangular distributions where, for $\alpha > 0$,

$$Tr(e,a,b) = \begin{cases} \frac{4(x-a)}{(b-a)^2} & a \leq x \leq (b+a)/2 \\ \frac{4(b-x)}{(b-a)^2} & (b+a)/2 < x \leq b \end{cases} \tag{1}$$

This is a legitimate family for the error E (see Appendix .2). Figure 1 shows the variance and Rényi's entropy of E plotted as functions of α. We observe that the variance has a minimum at $\alpha = 1$ while entropy attains its minimum value for $\alpha \to 0$ or $\alpha \to +\infty$, that is when the family converges to a Dirac function at zero, the optimal error solution.



(a) (b)

Figure 1: (a) Variance as a function of α. (b) Rényi's entropy as a function of α.

.2 Deriving the Error Distribution

$$\begin{aligned} F_E(e) &= P(E \leq e) = P((T=1, E \leq e) \vee (T=-1, E \leq e)) \\ &= P(T=1)P(E \leq e|T=1) + P(T=-1)P(E \leq e|T=-1) \\ &= \pi_1 P(1-Y \leq e|T=1) + \pi_{\text{-}1} P(-1-Y \leq e|T=-1) \\ &= \pi_1(1 - F_{Y|1}(1-e)) + \pi_{\text{-}1}(1 - F_{Y|\text{-}1}(-1-e)) \\ &= 1 - \pi_1 F_{Y|1}(1-e) - \pi_{\text{-}1} F_{Y|\text{-}1}(-1-e). \end{aligned} \tag{2}$$

The probability function (for the discrete case) and the density function (for the continuous case) are then easily obtained as

$$p_E(e) = \pi_1 p_{Y|1}(1-e)\delta(e, t-y) + \pi_{\text{-}1} p_{Y|\text{-}1}(-1-e)\delta(e, t-y) \tag{3}$$

$$f_E(e) = \frac{dF_E}{de} = \pi_1 f_{Y|1}(1-e) + \pi_{\text{-}1} f_{Y|\text{-}1}(-1-e) \tag{4}$$

respectively. It is worth noting that, for continuous class-conditional pdf's of Y we have $f_E(0) = 0$, since $\lim_{\epsilon \to 0} f_{Y|\text{-}1}(-1+\epsilon) = \lim_{\epsilon \to 0} f_{Y|1}(1-\epsilon) = 0$, as illustrated in Figure 2.

.3 Examples for Theorem 1

Example 1. We assume $\boldsymbol{\mu}_{\text{-}1} = (-5, 0)$, $\boldsymbol{\mu}_1 = (5, 0)$ and $\Sigma_1 = \Sigma_{\text{-}1} = I$. In this case $P_{\text{-}1} = P_1$ only if $p = 1/2$. The optimal decision line can be derived

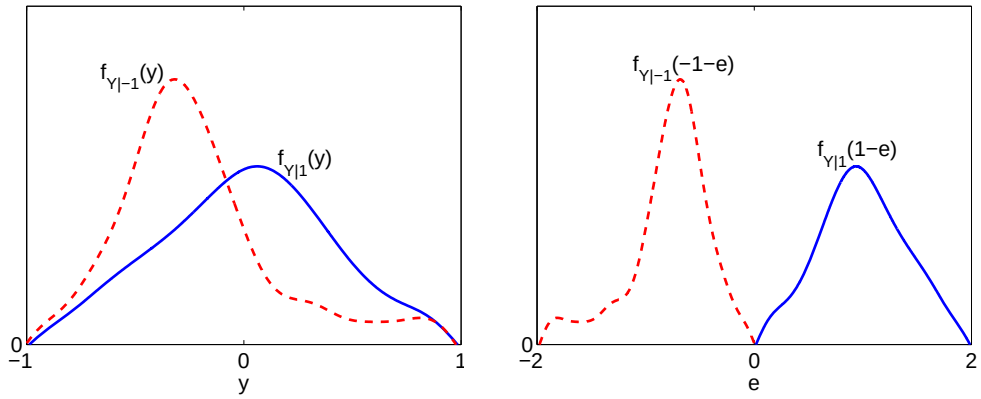


Figure 2: Illustration of the transformation $E = T - Y$, emphasizing the fact that $f_E(0) = 0$ for continuous class conditionals.

as

$$x_1^* = \frac{1}{10} \ln \left(\frac{1-p}{p} \right)$$

Hence,

$$-\frac{w_0^*}{w_1^*} = \frac{1}{10} \ln \left(\frac{1-p}{p} \right)$$

and therefore we can set

$$w_2^* = 0; \quad w_1^* = 1; \quad w_0^* = -\frac{1}{10} \ln \left(\frac{1-p}{p} \right)$$

Now, we can determine (numerically) that $\nabla H_S(\mathbf{w}^*) = \mathbf{0}$ only if $p = 1/2$; but this is the case of equal class error probabilities.

Example 2. In the second example we assume, $\boldsymbol{\mu}_{\text{-}1} = (-2, 0)$, $\boldsymbol{\mu}_1 = (2, 0)$, $p = 1/2$, $\Sigma_{\text{-}1} = I$ and

$$\Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

Although the covariance matrices are different the optimal decision line is still a vertical line with equation

$$x_1^* = -6 + \sqrt{32 + 2\ln(2)}$$

The error probabilities are unequal, with values $P_{\text{-}1} \approx 0.02$ and $P_1 \approx 0.03$.

We also verify that

$$\nabla H_S(1, 0, -6 + \sqrt{32 + 2\ln(2)}) \approx (-0.0153, 0, -0.0695) \neq \mathbf{0} \tag{5}$$

$$\nabla H_S(-1, 0, 6 - \sqrt{32 + 2\ln(2)}) \approx (0.0149, 0, 0.0672) \neq \mathbf{0} \tag{6}$$

$$\tag{7}$$

Thus, the optimal solution is indeed not a critical point of the error entropy.

Bibliography

[1] L. A. Alexandre and J. Marques de Sá. Error Entropy Minimization for LSTM Training. In *16th International Conference on Artificial Neural Networks*, 2006.

[2] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[3] D. Erdogmus, K. Hild II, and J. C. Príncipe. Blind Source Separation using Renyi's $\alpha-$marginal Entropies. *Neurocomputing*, 49:25–38, 2002.

[4] D. Erdogmus and J. C. Príncipe. Comparison of entropy and mean square error criteria in adaptive system training using higher order statistics. In *Intl. Conf. on ICA and Signal Separation*, pages 75–80, Helsinki, Finland, 2000.

[5] D. Erdogmus and J. C. Príncipe. An Error-Entropy Minimization Algorithm for Supervised Training of Nonlinear Adaptive Systems. *IEEE Transactions on Signal Processing*, 50(7):1780–1786, 2002.

[6] E. Gokcay and J. C. Príncipe. Information theoretic clustering. *IEEE Trans. on Pattern Analysis and Machine Learning*, 24(2):158–171, 2002.

[7] Long Han. *Kernel Partial Least Squares (K-PLS) for Scientific Data Mining*. PhD thesis, Rensselaer Polytechnic Institute, 2007.

[8] K. Hild II, D. Erdogmus, and J. C. Príncipe. Blind source separation using Renyi's mutual information. *IEEE Signal Processing Letters*, 8:174–176, 2001.

[9] R. Jenssen, K.E. Hild, D. Erdogmus, J. C. Príncipe, and T. Eltoft. Clustering using Rényis entropy. In *Int. Joint Conference on Neural Networks*, pages 523–528, 2003.

[10] R. Linsker. Self-organization in a perceptual network. *IEEE Computer*, 21:105–117, 1988.

[11] J. C. Príncipe, D. Xu, and J. Fisher. Information theoretic learning. In S. Haykin, editor, *Unsupervised Adaptive Filtering, vol. I: Blind Source Separation*, pages 265–319. Wiley, New York, 2000.

[12] M.D. Richard and R.P. Lippmann. Neural network classifiers estimate bayesian *a posteriori* probabilities. *Neural Computation*, 3:461–483, 1991.

[13] J. M. Santos, J. Marques de Sá, and L. A. Alexandre. Batch-sequential algorithm for neural networks trained with entropic criteria. In *International Conference on Artificial Neural Networks*, 2005.

[14] J. M. Santos, J. Marques de Sá, and L. A. Alexandre. Legclust - a clustering algorithm based on layered entropic subgraphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (accepted for publication)*, 2007.

[15] J. M. Santos, J. Marques de Sá, L. A. Alexandre, and F. Sereno. Optimization of the Error Entropy Minimization Algorithm for Neural Network Classification. In C. Dagli, A. Buczak, D. Enke, M. Embrechts, and O. Ersoy, editors, *Intelligent Engineering Systems through Artificial Neural Networks*, volume 14, pages 81–86. ASME Press Series, 2004.

[16] Jorge Santos. *Data classification with neural networks and entropic criteria*. PhD thesis, University of Porto, 2007.

[17] L. M. Silva, C. Felgueiras, L. A. Alexandre, and J. Marques de Sá. Error entropy in classification problems: A univariate data analysis. *Neural Computation*, 18(9):2036–2061, 2006.

[18] L. M. Silva, J. Marques de Sá, and L. A. Alexandre. Neural Network Classification using Shannon's Entropy. In *European Symposium on Artificial Neural Networks*, 2005.

[19] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26. Chapman & Hall, 1986.

[20] D. Stoller. Univariate two-population distribution free discrimination. *Journal of the American Statistical Association*, 49:770–777, 1954.