Neural Network Interest Group

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

Autor(es)/*Author(s)*:
Luís M. Silva

Relatório Técnico/*Technical Report* No. 3 /2007

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

Título/*Title*:

Data Classification with the Linear Perceptron and the EEM Principle I

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

Luís M. Silva

Relatório Técnico/*Technical Report* No. 3 /2007

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

Contents

Chapter 1

1.1 Minimizing error entropy

Consider the two-class problem where a given pattern $\mathbf{x} = (x_1, \ldots, x_d)^{\mathbf{t}}$ is to be classified in one of two classes, $\mathcal{C}_{\text{-}1}$ or $\mathcal{C}_1$. The linear discriminant is

$$y = \begin{cases} 1 & \sum_{i=1}^{d} w_i x_i + w_0 \geq 0 \\ -1 & \sum_{i=1}^{d} w_i x_i + w_0 < 0 \end{cases}, \tag{1.1}$$

where $w_i,\ i = 0, \ldots, d$ are real parameters. Geometrically, the decision surface, given by $\mathbf{w}^{\mathbf{t}}\mathbf{x} + w_0$ (where $\mathbf{w}^{\mathbf{t}} = (w_1, \ldots, w_d)$) is an hyperplane. A well known learning machine that implements this discriminant is the linear perceptron with Heaviside (threshold) activation function. We now study the data classification problem in light of the EEM principle. More precisely, we question if the solution (hyperplane) obtained by minimization of the error entropy corresponds to the optimal solution in the sense of minimum probability of error. The output Y in (1.1) and the true target T (class membership) can be viewed as discrete random variables with a given probability distribution. Thus we define the error random variable E by computing the difference of these variables

$$E = T - Y.$$

Obviously E is discrete and takes value in the $\{-2, 0, 2\}$ set with probabilities $P(E = -2) = P_{\text{-}1}$, the probability of misclassifying a pattern from class $\mathcal{C}_{\text{-}1}$, $P(E = 2) = P_1$, the probability of misclassifying a pattern from class $\mathcal{C}_1$ and $P(E = 0) = 1 - P_{\text{-}1} - P_1$, the probability of making a correct classification. The error entropy is then

$$H_E = -\left[P_{\text{-}1}\log(P_{\text{-}1}) + P_1\log(P_1) + (1 - P_{\text{-}1} - P_1)\log(1 - P_{\text{-}1} - P_1)\right]. \tag{1.2}$$

We now define, w.l.o.g., the following decision rule

$$\mathbf{x} \quad \text{belongs to} \quad \mathcal{C}_1 \quad \text{if} \quad \sum_{i=1}^{d} w_i x_i + w_0 \geq 0 \iff \mathbf{w^t x} + w_0 \geq 0.$$

Hence, we compute

$$P_{\text{-}1} = P(Y = 1, T = -1) = P(\mathbf{w^t x} + w_0 \geq 0, T = -1) = q(1 - F_{z|\text{-}1}(0)), \tag{1.3}$$

$$P_1 = P(Y = -1, T = 1) = P(\mathbf{w^t x} + w_0 \leq 0, T = 1) = pF_{z|1}(0), \tag{1.4}$$

where $F_{z|t}(0) = P(z \leq 0 | T = t)$ for $t \in \{-1, 1\}$; in other words, $F_{z|t}(0)$ is the conditional distribution value at the origin of the univariate random variable $z = \mathbf{w^t x} + w_0$. We proceed considering two cases:

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

$$wx + w_0 \geq 0 \iff x \geq -\frac{w_0}{w}. \tag{1.5}$$

   Surely, $w \neq 0$. Without loss of generality, we assume that class $\mathcal{C}_1$ is at the right of class $\mathcal{C}_{\text{-}1}$. 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 already studied [ref].

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

$$x_2 = -\left(\frac{w_1}{w_2}x_1 + \frac{w_0}{w_2}\right). \tag{1.6}$$

Note that cases (a) and (b) are quite similar to (1.5), 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.

1.2 Two Gaussian classes

We consider the two-class problem where the classes have bivariate Gaussian distributions. From the previous discussion, we see that it is crucial to determine the distribution of $z = \mathbf{w^t x} + w_0$. In this sense, we consider the following property of multivariate Gaussian distributions:

Property - Multivariate Gaussianity is preserved under linear transformations
If $\mathbf{x} = (x_1, \ldots, x_d)^\mathbf{t}$ has multivariate Gaussian distribution, $\mathbf{x} \sim G_d(\mathbf{m}, \Sigma)$, $\mathbf{w}_0 \in \mathbb{R}^m$ and $\mathbf{W}$ is a $m \times d$ real matrix, then:

$$\mathbf{z} = \mathbf{Wx} + \mathbf{w}_0 \sim G_m(\mathbf{Wm} + \mathbf{w}_0, \mathbf{W}\Sigma\mathbf{W^t}).$$

The variable $z = \mathbf{w^t x} + w_0$ above is the particular case with $d = 2, m = 1$. Thus, if $\mathbf{x} \sim G_2(\mathbf{m}, \Sigma)$, then:

$$z = \mathbf{w^t x} + w_0 \sim G_1(\mathbf{w^t m} + w_0, \mathbf{w^t}\Sigma\mathbf{w}).$$

We now consider two classes such that

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

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

$$F_{z|t}(0) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\sqrt{\mathbf{w^t}\Sigma_t\mathbf{w}}} \exp\left(-\frac{(x - \mathbf{w^t m}_t - w_0)^2}{2\mathbf{w^t}\Sigma_t\mathbf{w}}\right) dx \tag{1.7}$$

$$= \Phi\left(-\frac{\mathbf{w^t m}_t + w_0}{\sqrt{\mathbf{w^t}\Sigma_t\mathbf{w}}}\right) \tag{1.8}$$

and therefore

$$P_{\text{-}1} = q\left(1 - \Phi\left(-\frac{\mathbf{w^t m}_{\text{-}1} + w_0}{\sqrt{\mathbf{w^t}\Sigma_{\text{-}1}\mathbf{w}}}\right)\right), \tag{1.9}$$

$$P_1 = p\,\Phi\left(-\frac{\mathbf{w^t m}_1 + w_0}{\sqrt{\mathbf{w^t}\Sigma_1\mathbf{w}}}\right). \tag{1.10}$$

1.3 Graphical analysis

We are now ready to study H_E in some special cases. In order to make graphical representations we make the following assumptions:

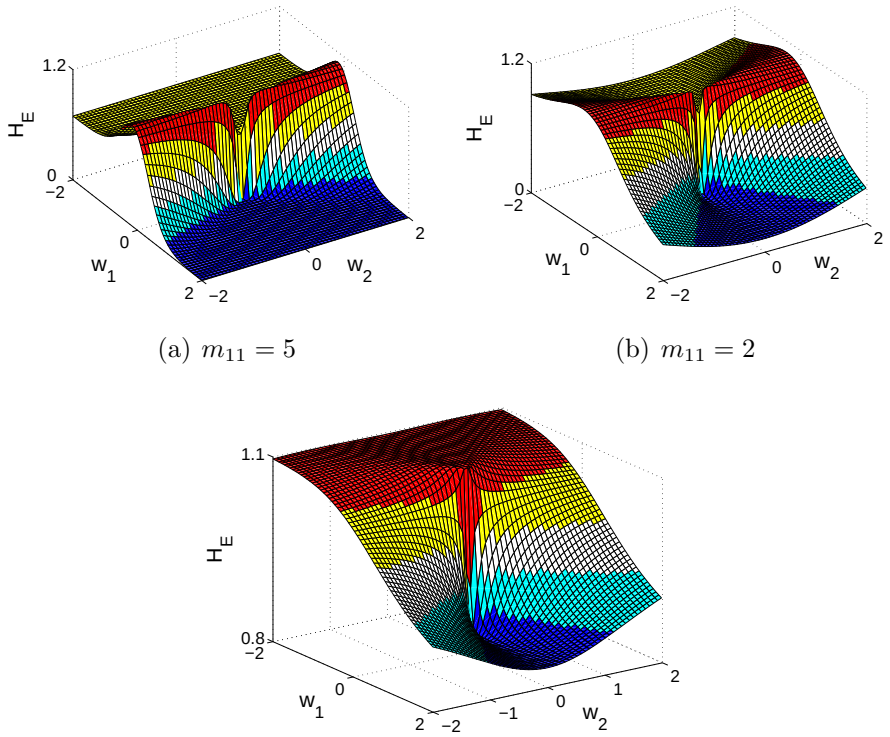
1. Considering $\mathbf{m}_t = (m_{t1}, m_{t2})$ for $t \in \{-1, 1\}$, we set $m_{t2} = 0$ and $m_{\text{-}11} = -m_{11}$ with $m_{11} > 0$. Basically, the centers of the classes lie in the horizontal axis and are symmetric to each other. Notice that every possible class configuration can be reduced to this case by applying shifts and rotations. As this does not alter the probabilities $P_{\text{-}1}$ and P_1, H_E 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[1] $\mathbf{w}^* = (w_1^*, w_2^*, w_0^*)^\mathbf{t}$, in the sense of minimum probability of error, is the vertical line $x_1 = 0$ and the optimal decision is to classify $\mathbf{x} = (x_1, x_2)^\mathbf{t}$ 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). We start, due to representational reasons, by setting $w_0 = 0$ and draw H_E as a function of w_1 and w_2 as shown in Figure 1.1. We first notice that the surfaces are symmetric about $w_2 = 0$. Let us analyze two different situations:

1. $w_1 > 0$
   As expected, the minimum of H_E is attained when $w_2 = 0$, independently of the proximity of the classes. For $w_2 \neq 0$ and considering sufficiently distant classes (Figure 1.1(a)), there are an infinite number of near-optimal solutions (the flat region) with nearly the same entropy. This corresponds to decision boundaries with high positive

[1] Henceforward, we will use this asterisk notation for the optimal solution.



(a) $m_{11} = 5$ (b) $m_{11} = 2$

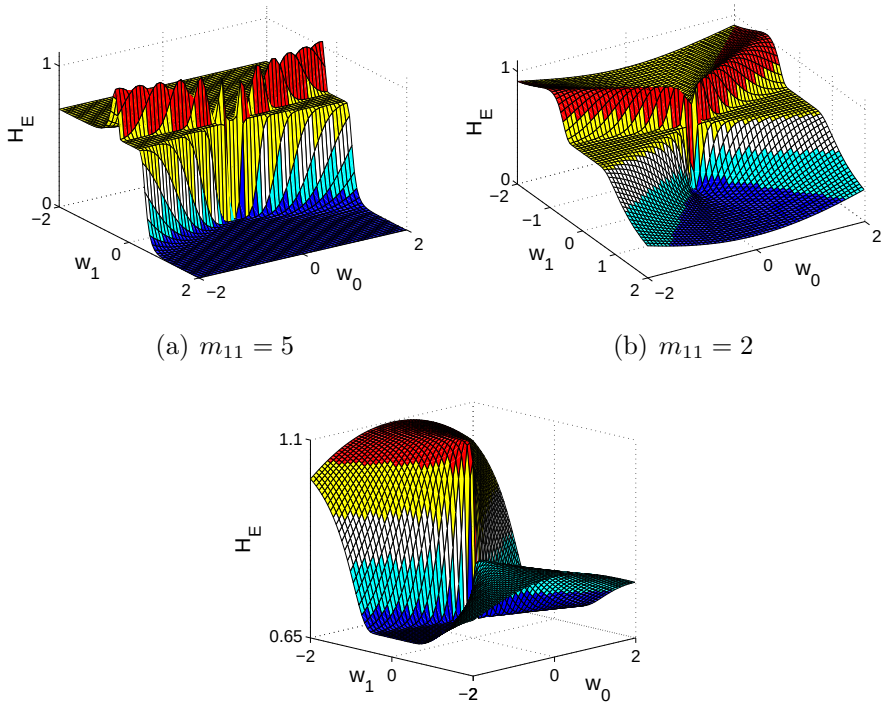
(c) $m_{11} = 0.5$

Figure 1.1: H_E for different values of $m_{11} = -m_{\text{-}11}$.

or negative slopes. Moreover, for smaller values of $\left|-\frac{w_1}{w_2}\right|$ (for example, fix w_2 and let $w_1 \to 0$), entropy increases, because we are considering solutions converging to $x_2 = 0$. When the classes are closer, the flat region disappears, because decision boundaries with slope are less tolerable here (there is more probability of error). We also notice that for distant classes, $H_E \approx 0$ for $w_2 = 0$ because $P(E = -2)$ and $P(E = 2)$ are nearly zero.

2. $w_1 < 0$
   In this case the perceptron is performing a swapped classification, $\mathcal{C}_{\text{-}1} \leftrightarrow \mathcal{C}_1$. Nevertheless, a minimum value is obtained for $w_2 = 0$, that seems to be preserved when the classes get closer in a similar fashion as for $w_1 > 0$.

These findings seem to contradict the previous results for Stoller splits. This is not true, however. Note that by setting $w_0 = 0$, the line is already fixed to the origin and only has the freedom to rotate. This is not the case for the Stoller split problem. Instead, if $w_2 = 0$, the line is fixed to be vertical and has freedom to make shifts – the Stoller split problem. Figure 1.2 shows similar plots for this case. The surfaces are very similar to the previous ones.

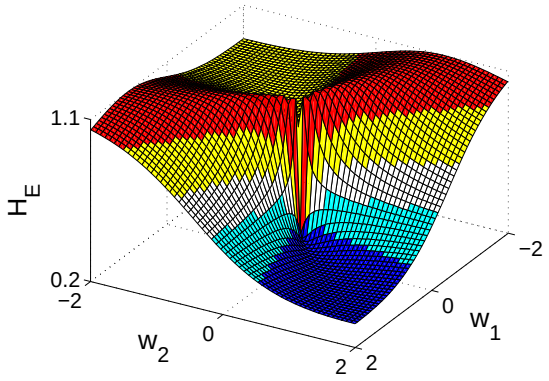


(a) $m_{11} = 5$ (b) $m_{11} = 2$

(c) $m_{11} = 0.5$

Figure 1.2: H_E for different values of m_{11} and $w_2 = 0$. The entropy maximum is clearly visible in (c) for $w_0 = 0$.

Again, we can distinguish, due to the same reasons, the case of $w_1 > 0$ and $w_1 < 0$. Let us analyze the former. When the classes are distant, the optimal solution is obtained when $w_0 = 0$, although small shifts of the line are also acceptable (the flat region). This is obvious because the probabilities P_1 and $P_{\text{-}1}$ are not greatly changed. However, when the classes get extremely close, the optimal solution, $w_0^* = 0$, turns out to be a maximum of entropy, which is in accordance to the results obtained for Stoller splits.
Finally, we illustrate a more general setting by assuming $\mathbf{m}_1 = -\mathbf{m}_{\text{-}1} =$



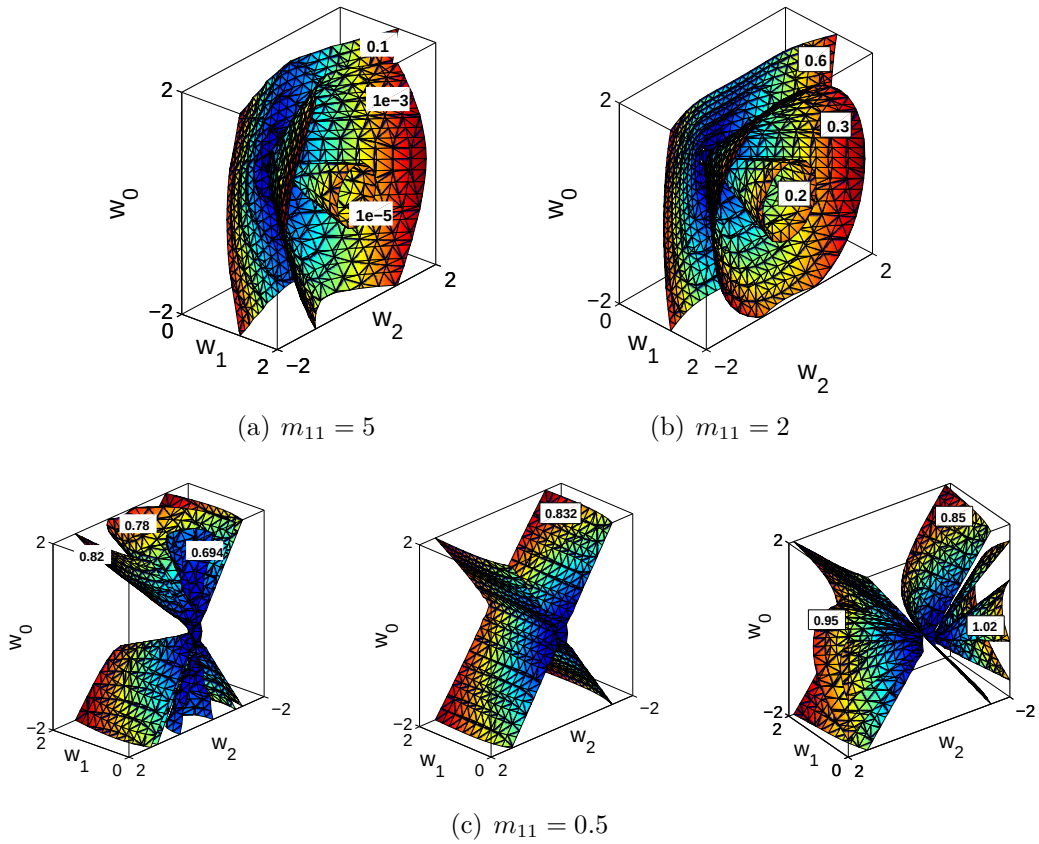
(a)

Figure 1.3: H_E for $\mathbf{m}_1 = -\mathbf{m}_{-1} = (1, 1)$ and $w_0 = 0$.

$(1, 1)^{\mathbf{t}}$ and noticing that the optimal decision line has equation $x_2 = -x_1$. From (1.6), we see that $w_0 = 0$ and $w_1 = w_2$. Again, for correct classification, w_1 (and hence w_2) must be positive. This is shown in Figure 1.3 (with $w_0 = 0$).

The previous analysis has shown different behaviors for $w_2 = 0$ and $w_0 = 0$. A natural question then arises: what is the behavior of H_E when all the parameters are free? More precisely, when training a learning machine that implements an hyperplane as the decision surface (like the simple perceptron), there is, in general, no a priori information that one or more of the parameters w_1, w_2 or w_0 should be zero. Thus, does the optimal set of parameters (in the sense of minimum probability of error) also correspond to a minimum of entropy? Does it turn to a maximum (like in the case of the Stoller split) when the classes get closer? We start investigating these questions by inspecting the surface levels of H_E, the equivalent to contour levels in the two variable case. In other words, we examine the surfaces $H_E(w_1, w_2, w_0) = c$ for increasing or decreasing values of $c \in \mathbb{R}$.

To draw the surfaces we assume the same conditions as above. Figure 1.4 shows some surface levels (iso-entropy surfaces or iso-entropics) of H_E. For distant classes (Figures 1.4(a) and 1.4(b)), we see that as we decrease the value of c, the iso-entropics converge to the positive w_1 axis, the optimal solution. On the other hand, when the classes are close ($m_{11} = 0.5$) the



(a) $m_{11} = 5$ (b) $m_{11} = 2$

(c) $m_{11} = 0.5$

Figure 1.4: Surface levels of H_E (with values of c represented). Figure (c) is split into three subfigures with increasing value of c from left to right.

behavior is completely different as is shown in Figure 1.4(c). Due to an high computational demand, this case is split into three subfigures. From left to right, we gradually increase the value of c. On the left, we find that when c is decreased to its minimum, the iso-entropics converge to the w_0 axis (although H_E is not defined there). On the other hand, when c is increased to its maximum (for $w_1 > 0$), the iso-entropics converge to the axis w_2; this is shown on the right figure of Figure 1.4(c). The optimal solution, the positive w_1 axis, appears for an intermediate value of c, as shown in the middle figure of Figure 1.4(c). In fact, we see from Figure 1.2(c) (although in this case $w_2 = 0$) that the positive w_1 axis is a local maximum.

1.4 First and second order information

Despite the above graphical suggestions, and specially in the last situation, we cannot conclude with confidence on the exact nature of these solutions. We now study these behaviors from an analytical point of view, using first and second order information about H_E (first and second order derivatives). In what follows we omit several expressions due to their complexity and length. It is easy to show that H_E is constant in the positive and negative axis w_1, with values converging to zero and $\ln(0.5)$ respectively, as the classes get farther. Also, by computing the gradient we see that $\bar{\mathbf{w}} = (w_1, 0, 0)^{\mathbf{t}}$ for $w_1 \neq 0$ are critical points of H_E or, in other words, that $\nabla H_E(\bar{\mathbf{w}}) = \mathbf{0}$. The nature of these critical points can be further investigated using second order information about H_E, which is given by its Hessian matrix. We restrict our study to the following cases:

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

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

   which is a positive semi-definite matrix (with $\det \nabla^2 H_E(\bar{\mathbf{w}}) = 0$). As 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_E to analyze its behavior in a neighborhood of $\bar{\mathbf{w}}$. Consider increments $\mathbf{h} = (h_1, h_2, h_3)^t$ where $h_i,\ i = 1, \ldots, 3$, is small. Then, we can write

$$H_E(\bar{\mathbf{w}} + \mathbf{h}) = H_E(\bar{\mathbf{w}}) + \mathbf{h}^{\mathbf{t}} \nabla H_E(\bar{\mathbf{w}}) + \mathbf{h}^{\mathbf{t}} \nabla^2 H_E(\bar{\mathbf{w}}) \mathbf{h} + o(\|\mathbf{h}\|^2). \quad (1.11)$$

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

$$H_E(\bar{\mathbf{w}} + \mathbf{h}) - H_E(\bar{\mathbf{w}}) \approx \mathbf{h}^{\mathbf{t}} \nabla^2 H_E(\bar{\mathbf{w}}) \mathbf{h}. \quad (1.12)$$

   Now, if the Hessian was positive definite (all positive eigenvalues), then for any $\mathbf{h}$, the quadratic form $\mathbf{h}^{\mathbf{t}} \nabla^2 H_E(\bar{\mathbf{w}}) \mathbf{h}$ 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 $\mathbf{h}^{\mathbf{t}} \nabla^2 H_E(\bar{\mathbf{w}}) \mathbf{h} = 0$; 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_E 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. $m_{11} = 0.5$ and $w_1 > 0$
   The Hessian now becomes

$$\nabla^2 H_E(\bar{\mathbf{w}}) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{0.2641}{w_1^2} & 0 \\ 0 & 0 & \frac{-0.1377}{w_1^2} \end{pmatrix}. \tag{1.13}$$

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

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

1.5 Minimum distance for Gaussian classes

It is worth asking what is the minimum distance between the Gaussian classes such that the minimum of H_E is preserved. This can be studied using the eigenvalues of the Hessian matrix. In fact, $\nabla^2 H_E(\bar{\mathbf{w}})$ is always a diagonal matrix with a zero entry, an always positive entry, and a third entry that changes the sign as the classes get closer (this is perfectly illustrated with the two Hessian matrices above); these entries are precisely the eigenvalues of the matrix. We can, therefore, determine the minimum distance to have a minimum of H_E at $(w_1, 0, 0)^{\mathbf{t}}$ for $w_1 \neq 0$, by inspecting when the last eigenvalue changes of sign. With $m_{11} = -m_{\text{-}11}$, $m_{12} = m_{\text{-}12} = 0$ and $\Sigma_1 = \Sigma_2 = \sigma^2 I$, the third eigenvalue can be written as a function of $d = m_{11}/\sigma$, which can be seen as a normalized half distance between the classes. The obtained expression is rather long and we just verify that the eigenvalue is positive if the following expression is positive:

$$\sqrt{2\pi} d (1 - \Phi(d)) \ln\left(\frac{2\Phi(d)}{1 - \Phi(d)}\right) - e^{-\frac{d^2}{2}}. \tag{1.14}$$

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.

1.6 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 bivariate version of Theorem 3 in [ref].

Theorem. In the bivariate two-class problem, if the optimal set of parameters $\mathbf{w}^* = (w_1^*, w_2^*, w_0^*)$ of a separating line constitute a critical point of 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, $\bar{z} = \mathbf{w^t x}$ is a projection of $\mathbf{x}$ onto $\mathbf{w}$. From an initial distribution represented by a density $g(\mathbf{x}) = qg_{\text{-}1}(\mathbf{x}) + pg_1(\mathbf{x})$ we get, on the projected space, the distribution of the projected data given by $f(\bar{z}) = qf_{\text{-}1}(\bar{z}) + pf_1(\bar{z})$. The parameter w_0 then functions 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 [ref], we can assert that $qf_{\text{-}1}(\bar{z}) = pf_1(\bar{z})$ at $\mathbf{w}^*$.

We rewrite the error probabilities of each class as

$$P_{\text{-}1} = q(1 - F_{\bar{z}|\text{-}1}(-w_0)), \tag{1.15}$$
$$P_1 = pF_{\bar{z}|1}(-w_0) \tag{1.16}$$

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

$$\frac{\partial P_{\text{-}1}}{\partial w_0} = -qf_{\bar{z}|\text{-}1}(-w_0) \tag{1.17}$$
$$\frac{\partial P_1}{\partial w_0} = pf_{\bar{z}|1}(-w_0) \tag{1.18}$$

From (1.2)

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

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

$$\frac{\partial H_E}{\partial w_0}(\mathbf{w}^*) = 0 \Leftrightarrow \tag{1.19}$$

$$\Leftrightarrow pf_{\bar{z}|1}(w_0^*)\left(\ln\left(\frac{1-P_{\text{-}1}-P_1}{P_{\text{-}1}}\right) - \ln\left(\frac{1-P_{\text{-}1}-P_1}{P_1}\right)\right) = 0 \tag{1.20}$$

$$\Leftrightarrow f_{\bar{z}|1}(w_0^*) = 0 \vee P_{\text{-}1} = P_1 \tag{1.21}$$

Note that $f_{\bar{z}|1}(w_0^*) = 0$ iff the classes have distribution with exclusive support (they are separable). But in this case $P_{\text{-}1} = P_1 = 0$. Thus, it remains that $P_{\text{-}1} = P_1$ is necessary to make null the above partial derivative.

q.e.d.

Computing the remaining partial derivatives with respect to w_1 and w_2 could show that the equal error probabilities condition is also sufficient. Thus, at this time, we do not know if there are situations such that at $\mathbf{w}^*$ with equal probabilities, the lacking derivatives (which for now, are difficult to compute) are null or not.

We illustrate the preceding Theorem, with the following two examples:

Example 1. We assume $\mathbf{m}_{\text{-}1} = (-5, 0)$, $\mathbf{m}_1 = (5, 0)$ and $\Sigma_1 = \Sigma_2 = I$. Also, $q = 1 - p$. Note that $P_{\text{-}1} = P_1$ only if $p = 1/2$. The optimal decision line becomes

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

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 $\nabla H_E(\mathbf{w}^*) = \mathbf{0}$ only if $p = 1/2$; but this is the case of equal class error probabilities.

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

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

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

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

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

$$\nabla H_E(1, 0, -6 + \sqrt{32 + 2\ln(2)}) \approx (-0.0153, 0, -0.0695) \neq \mathbf{0} \tag{1.22}$$

$$\nabla H_E(-1, 0, 6 - \sqrt{32 + 2\ln(2)}) \approx (0.0149, 0, 0.0672) \neq \mathbf{0} \tag{1.23}$$

$$\tag{1.24}$$

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