Neural Network Interest Group

Título/*Title*:
MSE, CE and MEE

Autor(es)/*Author(s)*:
J.P. Marques de Sá

Relatório Técnico/*Technical Report* No. 5 /2008

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

Título/*Title*:

MSE, CE and MEE

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

J.P. Marques de Sá

Relatório Técnico/*Technical Report* No. 5 /2008

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

MSE, CE and MEE

JP Marques de Sá
INEB, March 2008

1 Risk Functional

The risk functional of a machine provides a measure of how close the machine output $y = \varphi(x, w)$ is to the target variable t :

$$R(w) = \int_{X,T} L(t, \varphi(x; w)) dF(x,t)$$

The risk functional is the expected value of a loss function L on the set of all possible (x, t) pairs. The formula above is expressed as a Lebesgue integral in order to cope with any general probability distribution of (X,T). For instance, in Figure 1 we have a probability distribution which has two probability masses of p_1 and p_2 (two Dirac functions with those areas) added to a continuous component responsible by a probability of $1 - (p_1 + p_2)$. It doesn't make sense to speak of $f(x)$dx and express a probability calculation as a Riemann integral. However, it does make sense to speak of $dF(x)$ and probability calculations are then carried out as Lebesgue integrals.

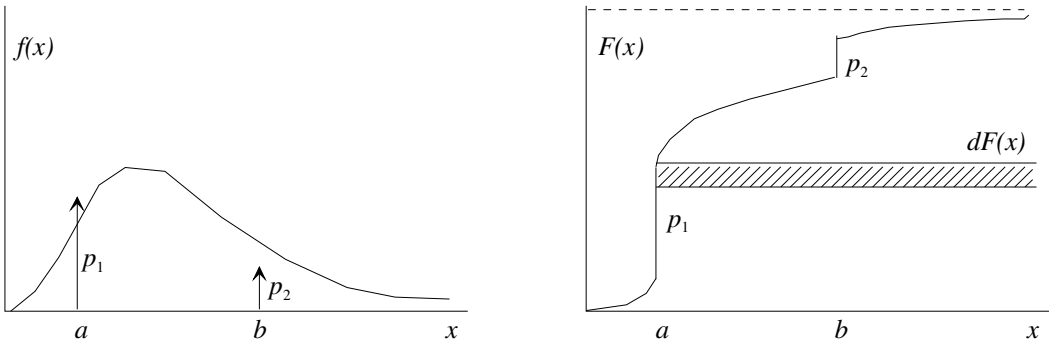


Figure 1

We now consider the data classification setting. The risk functional is rewritten as:

$$R(w) = \sum_{T} P(t) \int_{X} L(t, \varphi(x; w)) dF(x \mid t)$$

Furthermore, if we limit ourselves to the usual scenario of continuous data distributions (not mixed distributions as in Figure 1), we have:

$$R(w) = \sum_T P(t) \int_X L(t, \varphi(x; w)) f_X(x \mid t) dx$$

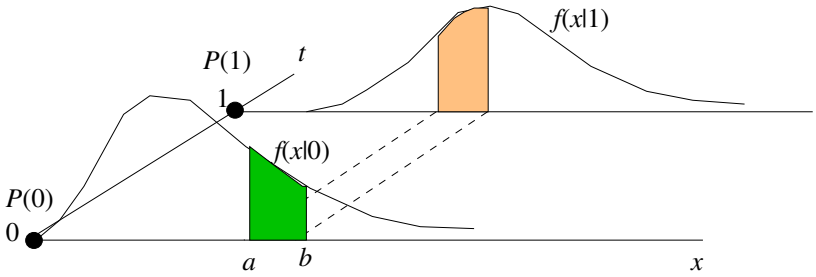


Figure 2. $E[g(x)]$ with $g(x) = 1$, $x \in [a, b]$, 0, otherwise, is green×$P(0)$ + red×$P(1)$.

The risk functional can be expressed in terms of any other variables by performing the adequate change of variable. In the following we assume a strictly increasing $\varphi(X)$ function (as is usual with sigmoids) and drop the dependence on w.

- The risk functional expressed in terms of the output variable:

$$f_Y(y \mid t) = \frac{f_X(x \mid t)}{\varphi'(x)} \quad (\varphi'(x) > 0) \text{ and } dy = \varphi'(x) dx \text{. Therefore:}$$

$$R(w) = \sum_T P(t) \int_Y L(t, y) f_Y(y \mid t) dy$$

$(Y = \varphi(X)$; see also Papoulis, pag. 142)

- The risk functional expressed in terms of the error variable:

For simplicity and convenience we will restrict ourselves to 2-class problems, $T = \{-1, 1\}$, $Y = \varphi(X) \in [-1,1]$. However, the results are directly generalized to other settings.

$$R(w) = P(-1) \int_{-1}^{1} L(t, y) f_Y(y \mid -1) dy + P(1) \int_{-1}^{1} L(t, y) f_Y(y \mid 1) dy$$

But, the error r.v. $E = T - Y$. For class -1 we have:

$$f_E(e \mid -1) = \frac{f_Y(-1 - e \mid -1)}{|-1|} = f_Y(y \mid -1); \; dy = -de$$

Therefore:

$$\int_{-1}^{1} L(t, y) f_Y(y \mid -1) dy = -\int_{0}^{-2} L(t, y) f_E(e \mid -1) de = \int_{-2}^{0} L(t, y) f_E(e \mid -1) de$$

Working out the second integral in the same way, we arrive at:

$$R(w) = P(-1)\int_{-2}^{0} L(t,y) f_E(e \mid -1)de + P(1)\int_{0}^{2} L(t,y) f_E(e \mid 1)de$$

In the case that $L(t,y) = (t-y)^2 = e^2$ we get:

$$R(w) = P(-1)\int_{-2}^{0} e^2 f_E(e \mid -1)de + P(1)\int_{0}^{2} e^2 f_E(e \mid 1)de = E_{T,E}\left[e^2\right]$$

This corresponds to the variance of the error only when $\mu_E = 0$.

Let us now take the following cost function:

$$L(t,y) = \begin{cases} 1 & \text{sign}(y) \neq t \\ 0 & \text{sign}(y) = t \end{cases} \qquad \text{with} \quad sign(y) = \begin{cases} 1 & y \geq 0 \\ -1 & y < 0 \end{cases}$$

This is the cost function corresponding to misclassifications with *soft* (e.g. sigmoidal) output: if $y \geq 0$ then decide class 1 else decide class 2.

Noting that $L(t, y) = 1$ if $y \in [0,1] \wedge t = 1 \vee y \in [-1,0] \wedge t = -1$, we have for this cost function:

$$P_e = R(w) = P(-1)\int_{-2}^{-1} f_E(e \mid -1)de + P(1)\int_{1}^{2} f_E(e \mid 1)de\,.$$

This affords an easy way to compute P_e (fig. 3).

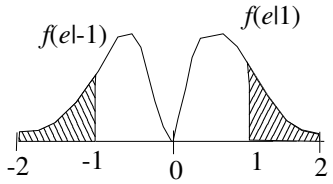


Figure 3. The dashed area corresponds to the error probability.

Note that in order to obtain zero error probability (separable classes for the function family implemented by the classifier) the error pdf most be confined to the [-1, 1] interval.
The cost function corresponding to misclassifications with *hard thresholded* output is:

$$L(t,y) = \begin{cases} 1 & \text{y} \neq t \\ 0 & y = t \end{cases}$$

In this case one may view the problem as a discrete error problem (the errors can only assume three values) and $P_e = 0$ corresponds to a discrete Dirac for the error (mass) probability function.

2 Entropy of Partitioned pdfs

Consider a pdf $f(x)$ which is a weighted sum of functions with disjoint domains, i.e.:

$$f(x) = \sum_i a_i f_i(x) \text{ , such that:}$$

a) Each $f_i(x)$ is a pdf defined in D_i
b) $\forall i, j \quad D_i \cap D_j = \varnothing$
c) f is defined in $D = \cup_i D_i$
d) $\sum_i a_i = 1$

We call such an $f(x)$ a *partitioned pdf*.

Then: $H_f = \sum_i a_i H_{f_i} - \sum_i a_i \ln(a_i)$

Proof:

$$H_f = -\int_D f \ln f = -\sum_i \int_{D_i} a_i f_i \ln(a_i f_i) = -\sum_i a_i \ln(a_i) \int_{D_i} f_i - \sum_i a_i \int_{D_i} f_i \ln(f_i)$$

- Note that entropy is invariant to scale reflections and translations:

$$H_{fY} = H_{f-Y+a}$$

A note concerning pdf and its moments (see Allan Gut, pag.158, 160, 175)

Definition: The characteristic function of the r.v. X is

$$\varphi_X(t) = E\left[e^{itx}\right] = \int_{-\infty}^{\infty} e^{itx} d_X F(x)$$

In the continuous case the characteristic function is similar to the Fourier transform of $f(x)$.

Theorem (uniqueness): Let X and Y be r.v. If $\varphi_X = \varphi_Y$ then $X = Y$ and conversely.

Theorem: Let X be a r.v. with distribution function F and characteristic function φ. If $E\left[|X|^n\right] < \infty$ for all n, and $\frac{|t|^n}{n!} E\left[|X|^n\right] \underset{n\to\infty}{\to} 0$ for all $t \in \Re$, then

$$\varphi(t) = 1 + \sum_{k=1}^{\infty} \frac{(it)^k}{k!} E\left[X^k\right]$$

3 Variance of Partitioned pdfs

Let $\mu = E_f[X]$; $\mu_i = E_{f_i}[X]$. We have:

$$V_f \equiv V[f] = \int_D (x-\mu)^2 f(x)dx = \sum_i a_i \int_{D_i} (x-\mu)^2 f_i = \sum_i a_i \int_{D_i} [(x-\mu_i) + (\mu_i - \mu)]^2 f_i$$

$$= \sum_i a_i V_{f_i} + \sum_i a_i (\mu_i - \mu)^2$$

Since $\int_{D_i} (x-\mu_i)(\mu_i - \mu) f_i = \mu_i^2 - \mu\mu_i - \mu_i^2 + \mu_i\mu = 0$

- Note that the variance is invariant to scale reflections but *not* invariant to scale translations.
- The mean is not invariant to scale reflections and translations.

Suppose that $f_{Y|-1}$ and $f_{Y||}$ are mutually symmetric, $p = q = ½$, we are given $V_{f_{Y|-1}}[Y]$ and want to compute $V_{f_E}[E]$. We first note that $V_{f_{Y|-1}}[Y] = V_{f_{Y||}}[Y] = V_{f_{-Y|-1}}[-Y]$, $\mu_{-1-Y|-1} = -1 - \mu_{Y|-1}$ and $\mu_E = 0$. Therefore, by the above result:

$$V_{f_E}[E] = 2\frac{1}{2} V_{f_{Y|-1}}[Y] + 2\frac{1}{2}(-1-\mu_{Y|-1})^2$$

4 Cross-Entropy

General Setting

We first consider a NN with c outputs y_k trained with n input vectors x_i.
Cross-entropy follows from the Kullback-Leibler (K-L) divergence of $p(t \mid x)$ in relation to $p_w(t \mid x)$:

$$KL(p \parallel p_w) = \int_X p(t \mid x) \ln \frac{p(t \mid x)}{p_w(t \mid x)} dx = E_p\left[\ln \frac{p(t \mid x)}{p_w(t \mid x)} \right]$$

That is, one is measuring the average "distance" between the unknown $p(t \mid x)$ and the machine produced $p_w(t \mid x)$, using as "distance" the log of the ratio.
The empirical estimate of the average given *n* inputs x_i is:

$$KL_n(p \parallel p_w) = \frac{1}{n} \sum_{i=1}^{n} \ln \frac{p(t_i \mid x_i)}{p_w(t_i \mid x_i)}$$

Assuming mutually exclusive classes and that the outputs $y_i(x_i)$ approximate the $p_w(t_i \mid x_i)$, we rewrite:

$$KL_n(p \parallel p_w) = \frac{1}{n} \sum_{i=1}^{n} \ln \frac{p_{1i}^{t_{1i}} \ldots p_{ci}^{t_{ci}}}{y_{1i}^{t_{1i}} \ldots y_{ci}^{t_{ci}}} = \frac{1}{n} \left(\sum_{i=1}^{n} \sum_{k=1}^{c} t_{ki} \ln(p_{ki}) - \sum_{i=1}^{n} \sum_{k=1}^{c} t_{ki} \ln(y_{ki}) \right)$$

We wish to train the NN such that $\min_W KL_n(p \parallel p_w)$ is reached. Since this minimum does not depend on the p_{ki}, one only has to minimize the following cross-entropy expression:

$$CE = -\sum_{i=1}^{n} \sum_{k=1}^{c} t_{ki} \ln(y_{ki})$$

In the same way as the K-L divergence, *CE* can be viewed as providing an empirical estimate of an average, corresponding to the following risk functional:

$$R_W = \sum_{k=1}^{c} P(t_k) \int_{X,T} L(t, y(x)) dF(x \mid t_k),$$

with

$$L(t, y) = -t \ln(y),$$

since

$$\hat{R}_W = -\sum_{k=1}^{c} \frac{n_k}{n} \left[\frac{1}{n_k} \sum_{i=1}^{n} t_{ki} \ln(y_{ki}) \right] = \frac{CE}{n}$$

Two-class setting

$$CE = -\sum_{i=1}^{n} t_{1i} \ln(y_{1i}) - \sum_{i=1}^{n} t_{2i} \ln(y_{2i})$$

Furthermore, let us consider a single output $y \equiv y_1$, $y \in D \subset [0,1]$, such that $t \equiv t_1 \in \{1\}$, $t_0 = 1 - t$; therefore, $1 - y$ represents y_0. We then have:

$$CE = CE_0 + CE_1 = -\sum_{i=1}^{n} t_i \ln(y_i) - \sum_{i=1}^{n} (1-t_i)\ln(1-y_i) = -\sum_{\substack{i=1 \\ x_i \in \omega_0}}^{n} \ln(1-y_i) - \sum_{\substack{i=1 \\ x_i \in \omega_1}}^{n} \ln(y_i)$$

The maximum support of y is: $D_{\max} = [0,1[_{\omega_0} \cup]0,1]_{\omega_1}$

CE is n times the empirical estimate of

$$R_W = -P(0)\int_{Y,T=0} \ln(1-y)dF(y \mid t=0) - P(1)\int_{Y,T=1} \ln(y)dF(y \mid t=1)$$

$$R_W = -P(0)E[\ln(1-y) \mid t=0] - P(1)E[\ln(y) \mid t=1]$$

Example 1
Let us consider $P(0) = P(1) = ½$ and the following family of uniform output pdf's:

$$f(y) = \begin{cases} u(0,d), & t=0 \\ u(1-d,1), & t=1 \end{cases}$$

We have:

$$CE = -2n\left[\frac{1}{2}E[\ln(y) \mid t=1]\right]$$

But:

$$E[\ln(y) \mid t=1] = \int_{1-d}^{1} \frac{1}{d}\ln(y)dy = \frac{1}{d}[y\ln(y) - y]_{1-d}^{1} = \frac{1}{d}\left(-(1-d)\ln(1-d) - d\right)$$

Therefore:

$$CE = \frac{n}{d}\left((1-d)\ln(1-d) + d\right)$$

Two particular cases:

1) $d = 0$, i.e., $f(y) = \begin{cases} \delta(y), & t=0 \\ \delta(1-y), & t=1 \end{cases}$: $\quad CE \underset{d\to 0}{\to} \dfrac{n(-d\ln(1-d) - 1 + 1)}{1} = 0$

2) $d = 1-\varepsilon$: $\quad CE = \dfrac{n}{1-\varepsilon}\left(\varepsilon\ln(\varepsilon) + 1 - \varepsilon\right) \underset{\varepsilon\to 0}{\to} n$

Numerical simulation:

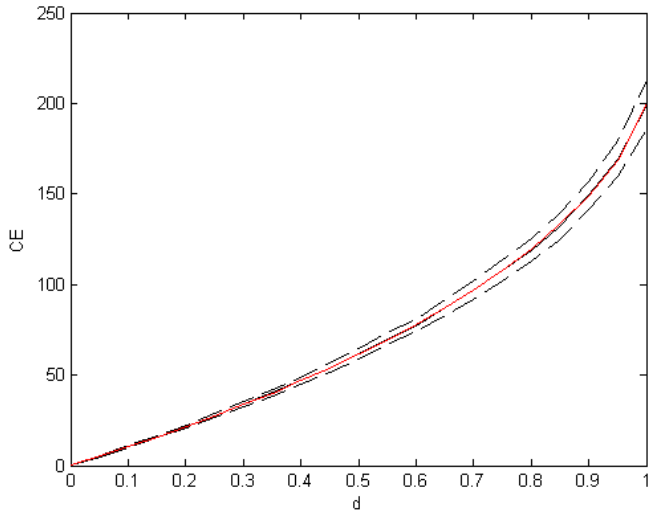


Figure 4. Average (solid black) ± std (dashed) of 50 experiments with n = 200 points, half belonging to each uniformly distributed class, for $d \in [0.0001, 0.999]$. The red curve is the theoretical curve. Note the convergence towards 0 and n.

We now search an expression of the cross-entropy in terms of the error r.v. For that purpose we first consider the expression of the 2-class cross-entropy with $T = \{-1, 1\}$, $Y = \varphi(X) \in [-1,1]$:

$$CE = -\sum_{i=1}^{n} \frac{1}{2}(1+t_i)\ln\frac{1+y_i}{1+t_i} - \sum_{i=1}^{n} \frac{1}{2}(1-t_i)\ln\frac{1-y_i}{1-t_i} = -\sum_{\substack{i=1 \\ t_i=-1}}^{n} \ln\frac{1-y_i}{2} - \sum_{\substack{i=1 \\ t_i=1}}^{n} \ln\frac{1+y_i}{2}$$

Which is n times the estimate of:

$$R = -P(-1)\int_{-1}^{1} \ln(1-y) f_Y(y \mid -1)dy - P(1)\int_{-1}^{1} \ln(1+y) f_Y(y \mid 1)dy + \ln 2$$

$$R = -P(-1)\int_{-2}^{0} \ln(2+e) f_E(e \mid -1)de - P(1)\int_{0}^{2} \ln(2-e) f_E(e \mid 1)de + \ln 2$$

Let us see the meaning of the two integrals. For the first one, we substitute $v_1 = 2 + e$, getting

$$\int_{-2}^{0} \ln(2+e) f_E(e \mid -1)de = E[\ln v_1 \mid -1]$$

Similarly, substituting $v_2 = 2 - e$ for the second integral:

$$\int_{0}^{2} \ln(2-e) f_E(e \mid 1)de = E[\ln v_2 \mid 1]$$

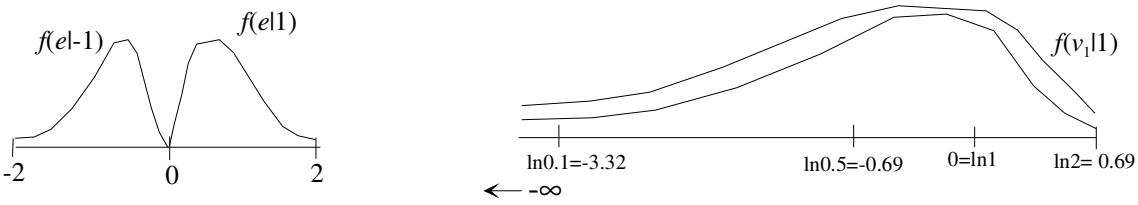


Figure 5

Therefore, as shown in Figure 13, the cross-entropy is ln2 minus the sum of the means of the logarithmically stretched conditional error pdfs.

If the conditional error pdfs are Dirac impulses at 0, we get:

$$E[\ln v_1 \mid -1] = E[\ln v_2 \mid 1] = \ln 2 \quad \Rightarrow \quad -P(-1)\ln 2 - P(1)\ln 2 + \ln 2 = 0$$

As the conditional error pdfs concentrate near the extremes the means $E[\ln v_1 \mid -1]$, $E[\ln v_2 \mid 1]$, become very large negative numbers and therefore CE becomes very high.

5 The Three Risk Functionals

For simplicity we consider $T = \{-1, 1\}$, $Y = \varphi(X) \in [-1,1]$:

$$R_{MSE}(e;w) = \sum_{t\in\{-1,1\}} P(t) \int_{t-1}^{t+1} e^2 \; f_E(e \mid t)\, de$$

$$R_{CE}(e;w) = \sum_{t\in\{-1,1\}} P(t) \int_{t-1}^{t+1} -\ln(2-te) \; f_E(e \mid t)\, de$$

$$R_H(e;w) = \sum_{t\in\{-1,1\}} P(t) \int_{t-1}^{t+1} -\ln f_E(e \mid t) \; f_E(e \mid t)\, de - \sum_{t\in\{-1,1\}} P(t)\ln P(t)$$

We now consider the *adaptive adjustment* of the risk functionals, i.e., that there is an adaptive algorithm that at each step tries to move towards the minimization of the risk functionals.
For the first two risk functionals at a certain step characterized by an error e, this corresponds to minimizing an average distance expressed as a function of e:

$$L_{MSE}(e) = e^2 = (t-y)^2$$
$$L_{CE}(e) = -\ln(2-te) = -\ln(2-t(t-y))$$

Example for $t = 1$:

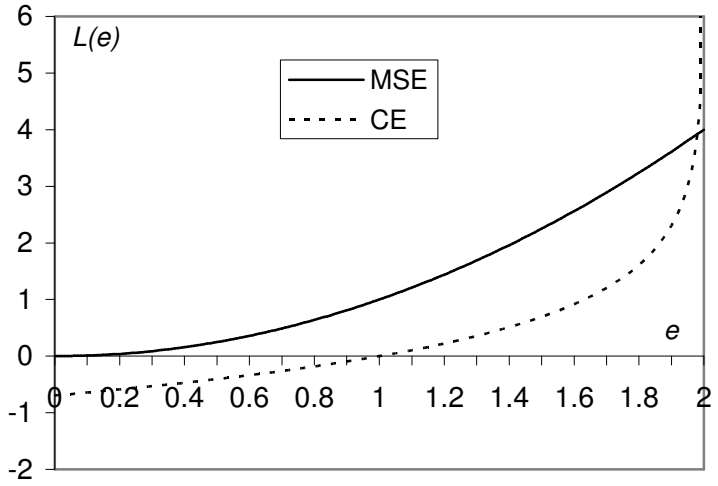


Figure 6

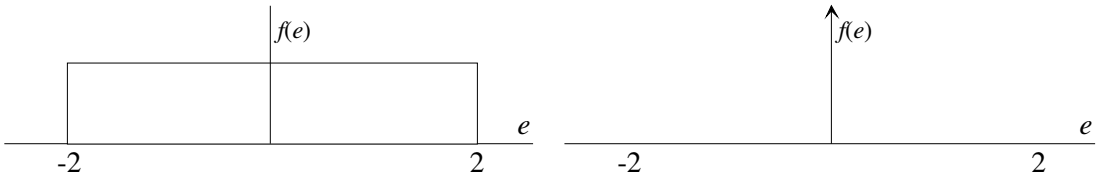
MSE provides a square dispersion measure (related to the error variance): MSE "doesn't like" long-tail distributions, and therefore outliers.

CE provides a logarithmic dispersion measure in such a way that true errors – note that for $e < 1$ the output y is not in error - are heavily weighted: CE penalizes heavier than MSE the tails corresponding to true errors and more so when the outputs are completely in the wrong side (true outliers). On the other hand, CE penalizes less than MSE the non-true errors.

The H risk functional is not based on a *constant* distance measure $L(e)$, since at any adaptive step the whole error distribution has to be taken into account. Thus, we are using a *variable* distance measure: $L(e;w) = -\ln f_E(e;w)$. When analyzing MEE one has to take into account *families* of error distributions.

We know that the uniform distribution is the maximum entropy distribution among all continuous distributions which are supported in an interval $[a, b]$. We also know that the minimum entropy distribution is the Dirac δ-function, which may be seen as the limit of a uniform family[1]. So, ideally we would like that even in the initial worst case of H we would converge to the Dirac function at zero:

[1] In order to comply with the continuity requirement of error pdfs, we may always suppose without any important change of the main results that the end jumps are made by arbitrarily high slope lines.



a) $H = \ln 4 = 1.386$ b) $H = -\infty$

Figure 7

1) The problem is that entropy is invariant to partitions and scale translations; so, from a) and during the minimizing process one could fall in equally "ordered" distributions but probably meaning very different things in terms of classifier performance:

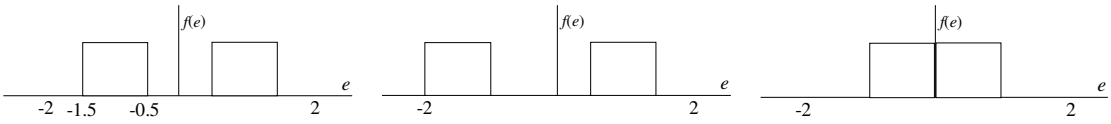
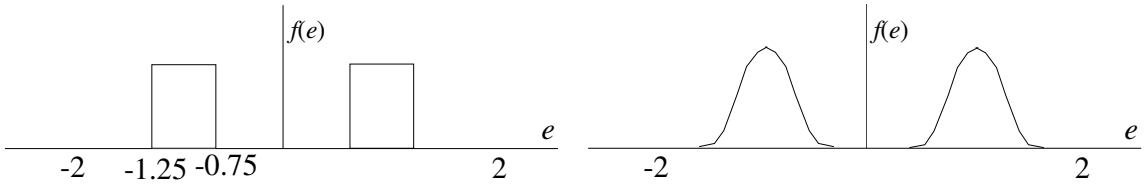


Figure 8. All these error distributions have $H = \ln 2 = 0.693$.

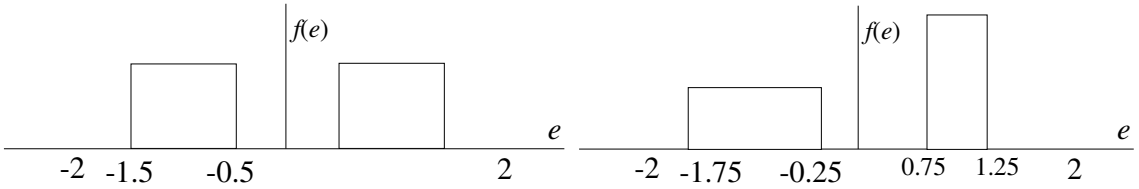
2) For equal variance of the errors MEE "prefers" completely ordered distributions (values concentrated in smaller intervals) than tailed distributions:



(b-$a = 0.5$) $H = 0$ ($\sigma = 0.14434$) $H = 0.176$

Figure 9

3) Entropy tolerates lack of order in one component if the other gains in order:



a) $H = 0.693$ b) $H = 0.549$

Figure 10. In both cases the intervals occupied by the distributions have a total measure of 2. In b) one of the intervals increased by 0.5 (decrease of order) and the other increased by the same amount (increase of order). However, the increase of entropy on the left component (0.2027) is more than compensated by the decrease on the right one (-0.3466).

As a consequence of this property MEE tolerates outlier errors as far as the bulk is very concentrated.

Properties 2 and 3 are clearly advantageous in many circumstances. Property 1 can bring about difficulties when applying MEE.

6 Discriminative Training with MSE, CE and MEE

We now consider given a training set and adaptively fitting a function $\phi_w \in \Phi$ in order to converge to the minimum value of the empirical risk functional.

We may, for instance, consider the adaptive process consist of gradient descent in the hypersurface corresponding to the empirical risk functional, i.e., gradient descent in $\hat{R}(\mathbf{w})$.

Using MSE one tries to minimize the variance of the errors (deviations).
Using CE one tries to minimize the logarithm of the error (deviation) amplitudes.
Using MEE one tries to move to the error (deviation) distribution with smaller entropy.

We assume $\hat{R}_{MEE}(\mathbf{w})$ computed from $\hat{f}_Y(y \mid t)$.

Let us see with the help of Figure 10, where we assume Φ as the straight line family, what can happen during the adaptive training.

In a) we have a possible initial situation: all the errors are at one side and widely spread $\Rightarrow$ MSE, CE and MEE are high.

Next, we move on the hypersurface in order to:

MSE: decrease the error variance.
CE: decrease the logarithm of the error amplitudes.
MEE: get $\hat{f}_Y(y \mid t)$ achieving smaller error entropy.

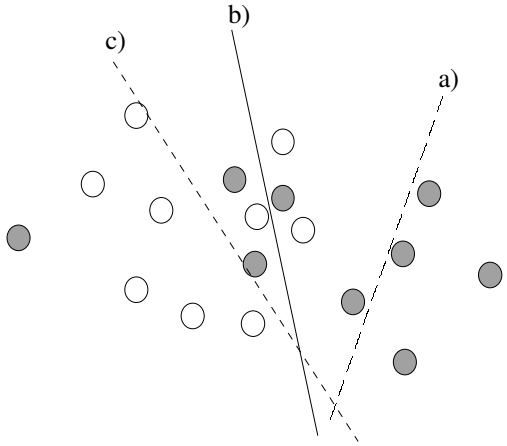
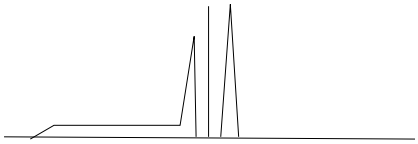
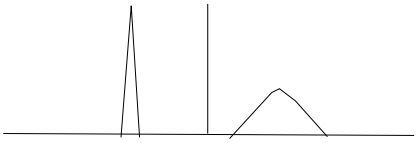


Figure 11

In b) we reach the best MEE: 

In c) we reach the best MSE: 

7 Linear Discriminant Output pdf

A linear discriminant (l.d.) transforms the input pdf's into an output pdf by orthogonal projection onto the weight vector.

We analyze for simplicity the 2D case. Consider the l.d. transformation:

$$d(\mathbf{x}) = w_1 x_1 + w_2 x_2 + w_0 = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + w_0 = \mathbf{w}'\mathbf{x} + w_0$$

where **w** is the weight vector.

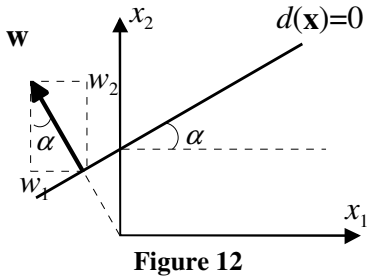


Figure 12

The decision border $d(\mathbf{x}) = 0$ is a straight line orthogonal to the weight vector. Let us confirm this:

$$d(\mathbf{x}) = w_1 x_1 + w_2 x_2 + w_0 = 0 \quad \Rightarrow \quad x_2 = -\frac{w_1}{w_2} x_1 - \frac{w_0}{w_2} \quad \Rightarrow \quad \tan(\alpha) = -\frac{w_1}{w_2}$$

On the other hand, a vector $\mathbf{n}$ orthogonal to the decision border with amplitude r and in the direction shown above has components:

$$n_1 = -r \sin\alpha = -r \frac{\tan(\alpha)}{\sqrt{1+\tan^2(\alpha)}} = r \frac{w_1}{\|\mathbf{w}\|}$$

$$n_1 = r \cos\alpha = r \frac{1}{\sqrt{1+\tan^2(\alpha)}} = r \frac{w_2}{\|\mathbf{w}\|}$$

This confirms the orthogonality of $\mathbf{w}$ to the decision border. Therefore, since $d(\mathbf{x})$ is the dot product of $\mathbf{w}$ and $\mathbf{x}$, in order to obtain the pdf of $d(\mathbf{x})$ one only has to project the data points onto the direction defined by $\mathbf{w}$.

We now analyze the case where the line passes through the origin and we want to express the projection $d(\mathbf{x}) = \mathbf{w}'\mathbf{x}$ in terms of the angle α that the line makes with the horizontal axis. Since $\tan(\alpha) = -w_1/w_2$ we have $w_1 = -\sin(\alpha)$, $w_2 = \cos(\alpha)$; therefore,

$$d(\mathbf{x}) = -\sin(\alpha) x + \cos(\alpha) y$$

8 Examples where MEE Does Not Always Solve the Classifier Problem

We consider 2-class classification problems in bivariate space $\Re^2$, target space $T = \{-1, 1\}$. We denote the input vectors by $[x_1\ x_2]'$ and consider the following marginal pdf's:

$$f_{-1}(x_1) = \frac{1}{2} n_{-}(1, \sigma, d), \ f_1(x_1) = f_{-1}(-x_1),$$

$$f_t(x_2) = u(-\frac{c}{2}, \frac{c}{2}), \ \text{with } c > 0 \text{ and}$$

where $n_{-}(\mu, \sigma, d)$ is the lower half of the Gaussian pdf with mean μ and std σ, truncated at d, and $u(a,b)$ is the uniform distribution in $[a,b]$.

Example 2

$$\sigma = 0.75, \ \mu = -1, \ d = 1, \ c = 1 \ \text{(dataset znormal200)}$$

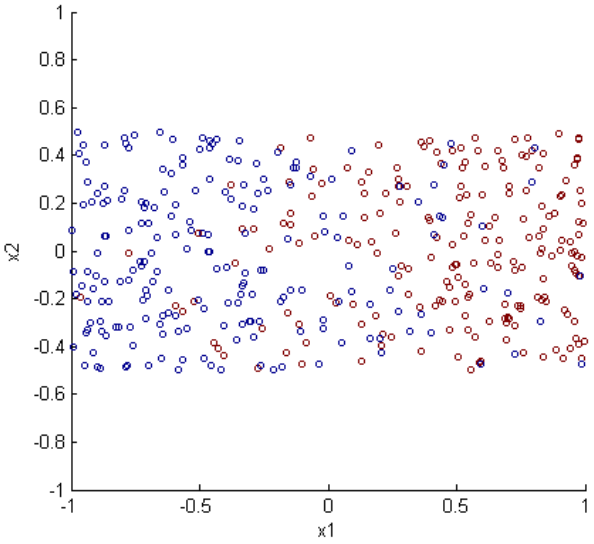


Figure 13. Example 2 dataset with 200 points per each class.

The estimated H, V were computed as follows (HVemp function):

1. Projection of all points over the **w** line (computation of $d(\mathbf{x})$) for $\alpha = -\pi/2, 0$.
2. Computation of the error values.
3. Computation of the variance of the errors.
4. Estimation of $\hat{f}_E(e)$ using a Gaussian kernel with $h = 0.1$ in a grid with $d_s = 0.05$ increments. (Both h and d_s were selected based on several experiments.)
5. Computation of H by integration.

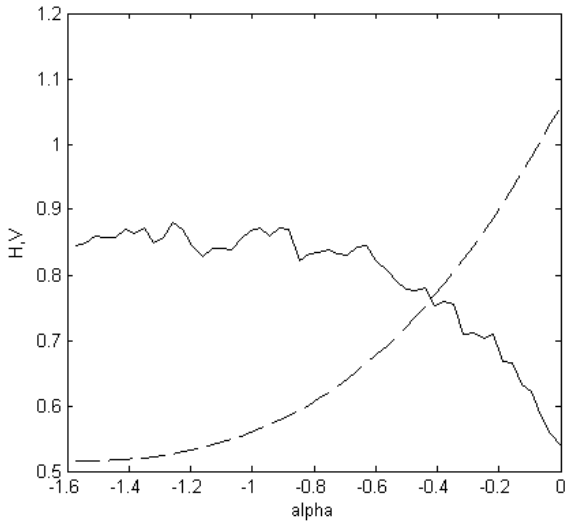


Figure 14. H,V for znormal200 dataset

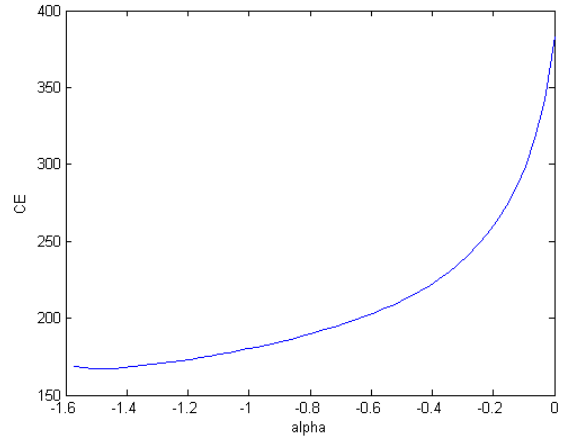


Figure 15. CE for znormal200 dataset

In this example, MEE fails to produce the good solution. The problem is that MEE looks at the most concentrated pdf, which happens to occur for $\alpha = 0$.

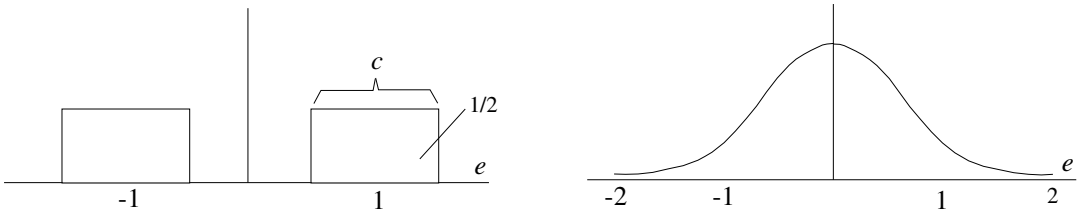


Figure 16

MEE will provide the good solution if either c is increased or σ decreased, as in the following examples.

Example 3

$$\sigma = 0.75,\ \mu = -1,\ d = 1,\ c = 2 \quad \text{(dataset znormal200-1)}$$

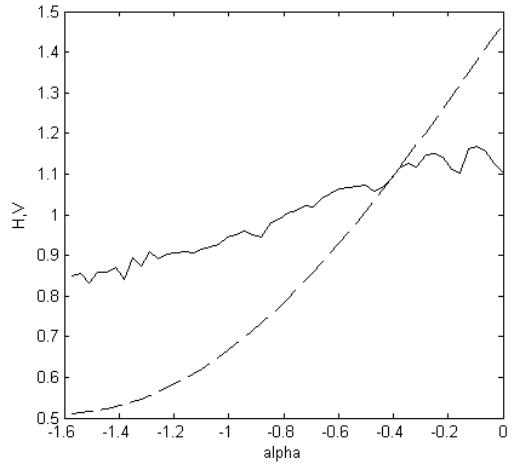


Figure 17

Example 4

$$\sigma = 0.4,\ \mu = -1,\ d = 1,\ c = 1 \quad \text{(dataset znormal200-s04)}$$

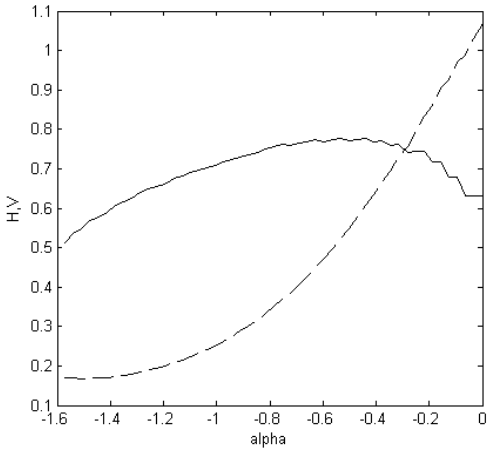


Figure 18

In both examples 3 and 4 CE produces the good solution (similar to Fig. 8).

9 Examples where MEE solves the Classifier Problem and MSE and CE Do Not

9.1 Setting

We consider 2-class classification problems in bivariate space $\Re^2$, target space $T = \{-1, 1\}$. We denote the input vectors by $[x_1\ x_2]'$ and consider the following marginal pdf's:

$$f_1(x_1) = \frac{1}{2}(u(a,1) + u(b,a)),\ f_{-1}(x_1) = f_1(-x_1),$$

$$f_t(x_2) = u(-\frac{c}{2}, \frac{c}{2}),\ \text{with } a, c > 0 \text{ and}$$

where $u(a,b)$ is the uniform distribution in $[a,b]$.

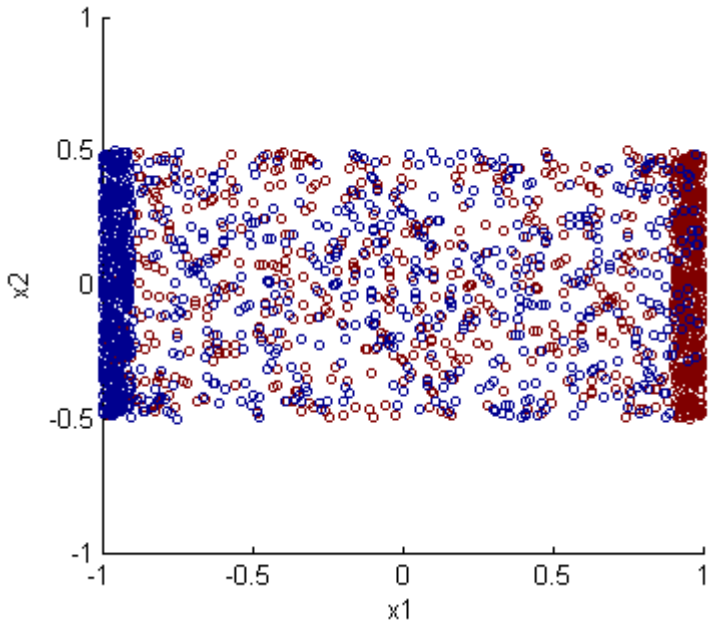


Figure 19. 1000 points of each class with $a = 0.9$, $b = -1$, $c = 0.5$ (500 points in each class rectangle).

Given a balanced training set of the two classes the classification problem consists of adjusting a straight line passing through the origin ($x_2 = \tan(\alpha)x_1$) yielding the minimum probability of error.

We first compute the theoretical MEE and MSE for two configurations, with $p = q = 1/2$:

Configuration #1: $\alpha = -\pi/2$; $\mathbf{w} = [0\ 1]'$; $d(x_2) = 0$

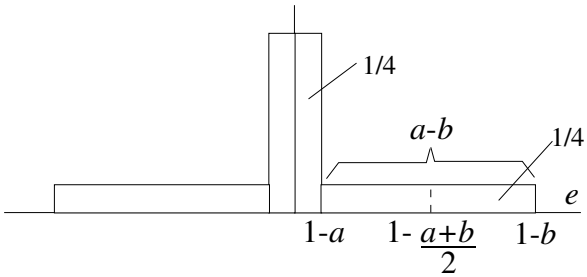


Figure 20. Configuration #1 represented in terms of the error variable ($E = T - D(X)$)

$$H = 2\left[\frac{1}{4}\ln(1-a) - \frac{1}{4}\ln\frac{1}{4}\right] + 2\left[\frac{1}{4}\ln(a-b) - \frac{1}{4}\ln\frac{1}{4}\right] = \frac{1}{2}\ln(1-a) + \frac{1}{2}\ln(a-b) - \ln\frac{1}{4}$$

$$V = \frac{1}{2}\frac{[2(1-a)]^2}{12} + 2\frac{1}{4}\frac{(a-b)^2}{12} + 2\frac{1}{4}\left[1 - \frac{a+b}{2}\right]^2$$

Configuration #2: $\alpha = 0$; $\mathbf{w} = [1\ 0]'$; $d(x_1) = 0$

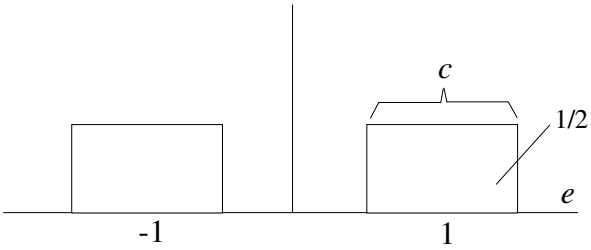


Figure 21. Configuration #2 represented in terms of the error variable ($E = T - D(X)$)

$$H = 2\left(\frac{1}{2}\ln c\right) - 2\left(\frac{1}{2}\ln\frac{1}{2}\right) = \ln c - \ln\frac{1}{2}$$

$$V = 2\left(\frac{1}{2}\frac{c^2}{12}\right) + 2\frac{1}{2}1^2 = \frac{c^2}{12} + 1$$

9.2 MSE and MEE

Example 5

$$a = 0.9,\ b = -1,\ c = 1.$$

We have (theoretical values):

$$H_{\#1} = 0.556;\ H_{\#2} = 0.693$$

On the other hand:

$$V_{\#1} = 0.703;\ V_{\#2} = 1.083$$

Using numerical simulation for several values of α ($-\pi/2{:}\pi/100{:}0$) we obtained the results of Figure 15.

(We assume the decision line rotating in $[-\pi/2, 0]$ in order not to swap the class labels. Note that for $\alpha = \pi/2$ we get $d(\mathbf{x}) = -x$ and class 1 would have values corresponding to class −1 and vice-versa. Although this swap is of no consequence for MSE and MEE it would have to be taken into account for CE.)

The numerical simulation (HV function) consisted of:

1. Generation of 4000 points of each class equally distributed between the two uniform "rectangles" (each rectangle weight is ½, i.e., gets 2000 points).
2. Projection of class −1 points over the **w** line (computation of $d(\mathbf{x})$).
3. Estimation of $f_{-1} \equiv f_{\mathbf{x}\in C-1}(d(\mathbf{x}))$ using a Gaussian kernel with $h = 0.006$ in a d grid with $d_s = 0.0025$ increments. (Both h and d_s were selected as the optimal values yielding the least sum of absolute deviations from the theoretical H and V values for $\alpha = -\pi/2, 0$).
4. Computation of H and V.

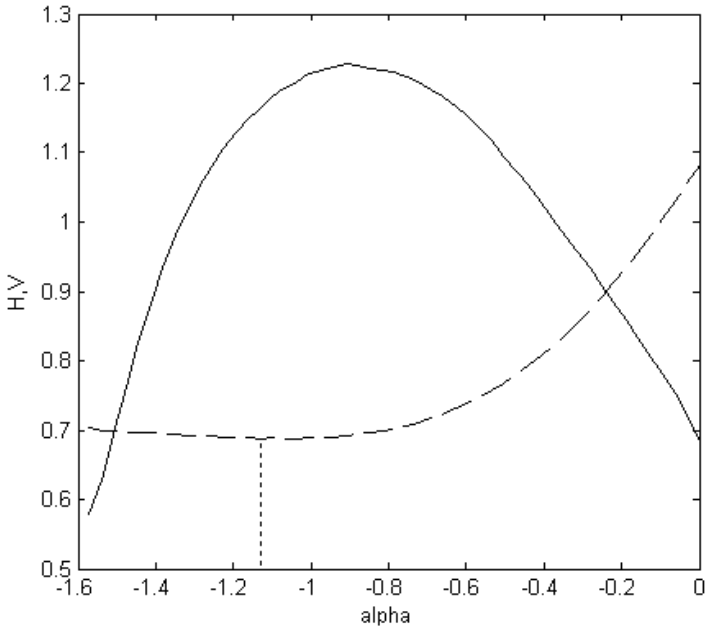


Figure 22 Entropy (solid line) and variance (dash line) as functions of the angle alpha (dataset z2000).

The MEE method achieves the smallest possible error at $\alpha = -\pi/2$:

Error = 0.263

The MSE method also achieves the smallest error at $\alpha = -1.13$:

Error = 0.263

So in this case, although the MSE decision line is not the "best" line, still the error is not affected.

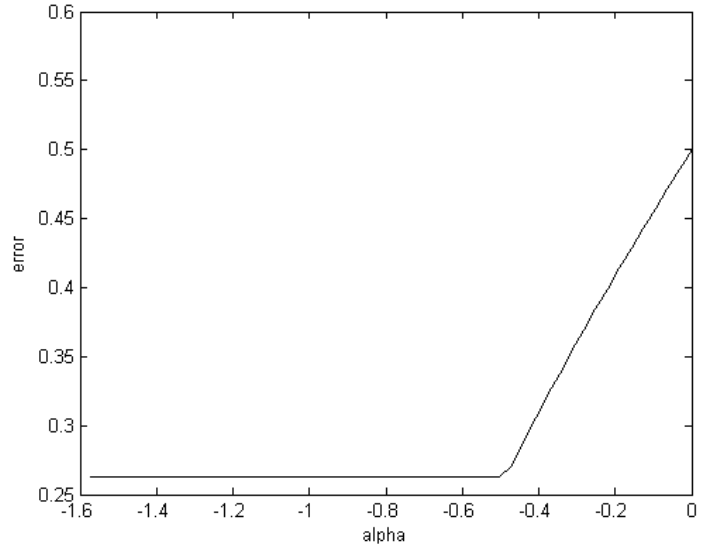


Figure 23. The error curve for Example 5.

Note that if we perform the computation of H, V using the previous "empirical estimate" procedure, the results obtained are essentially the same:

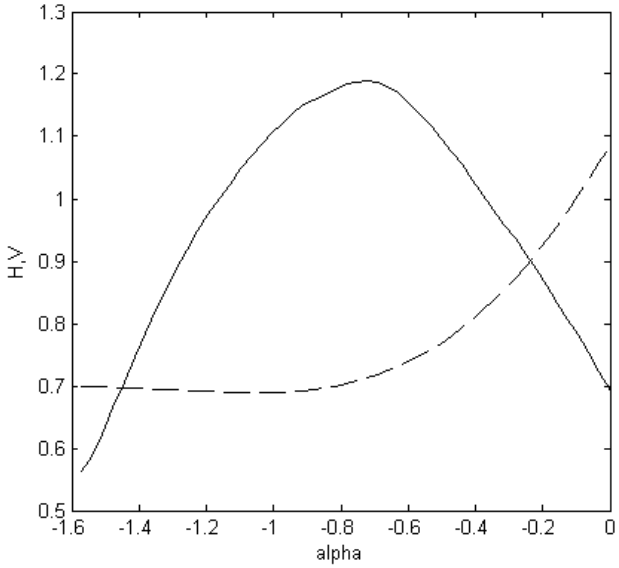


Figure 24

Example 6

$$a = 0.95,\ b = -1.7,\ c = 0.9.$$

We have:

$H_{\#1} = 0.378$; $H_{\#2} = 0.588$

$V_{\#1} = 1.238$; $V_{\#2} = 1.068$

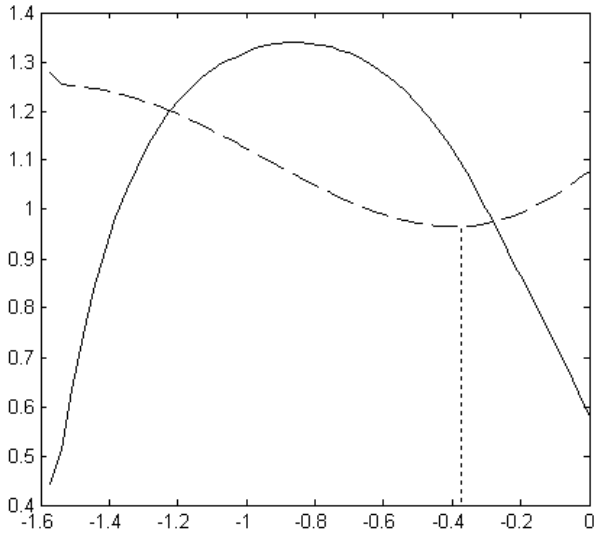


Figure 25 Entropy (solid line) and variance (dash line) as functions of the angle alpha (dataset v2000).

The MEE method achieves the smallest possible error at $\alpha = -\pi/2$:

Error = 0.321

The MSE method achieves the smallest error at $\alpha = -0.377$:

Error = 0.355

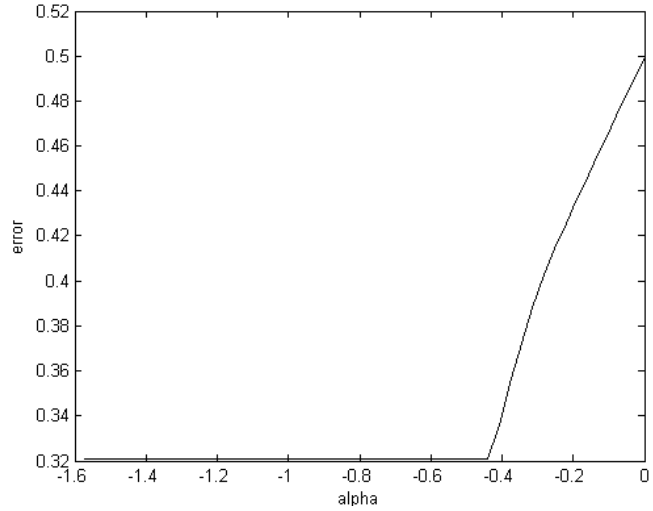


Figure 26. The error curve for Example 2

Even with a smaller number of points the same results can be arrived at (provided adequate h and d_s are used).

Figure 20 shows the results obtained when 500 points per class are used ($h = d_s = 0.02$; empirical estimate procedure).

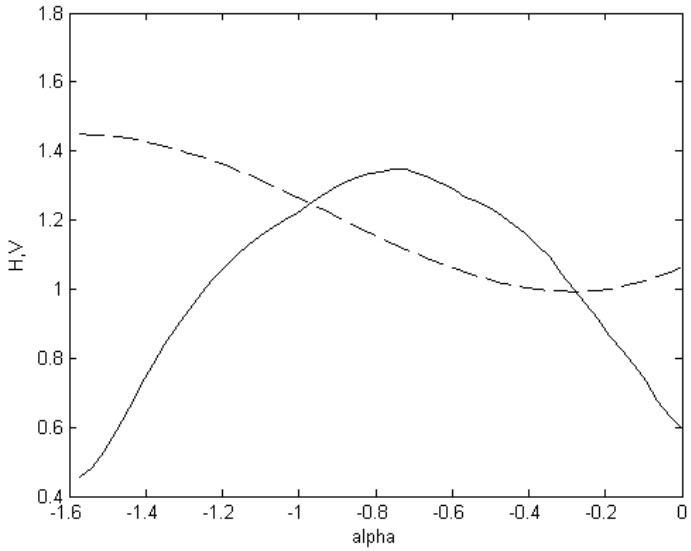


Figure 27

9.3 Cross-Entropy Results

Example 7
Same dataset as in Example 5.

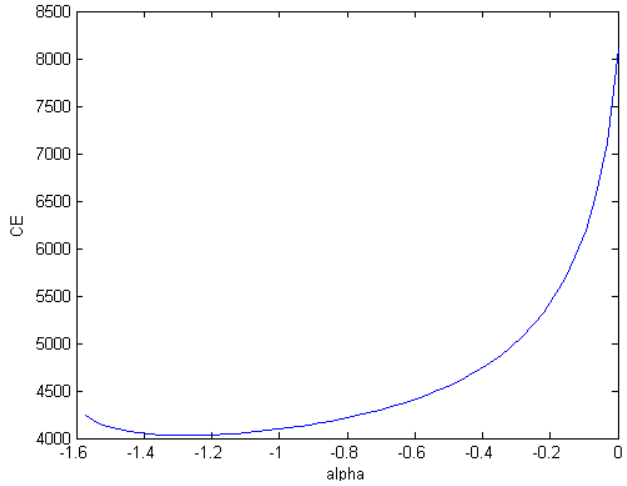


Figure 28. Cross-entropy for Example 5 dataset (z2000).

Example 8

$$a = 0.95,\ b = -2.4,\ c = 0.9.$$

In this example we only consider the theoretical error curves computed by numerical simulation as above (4000 points per class).

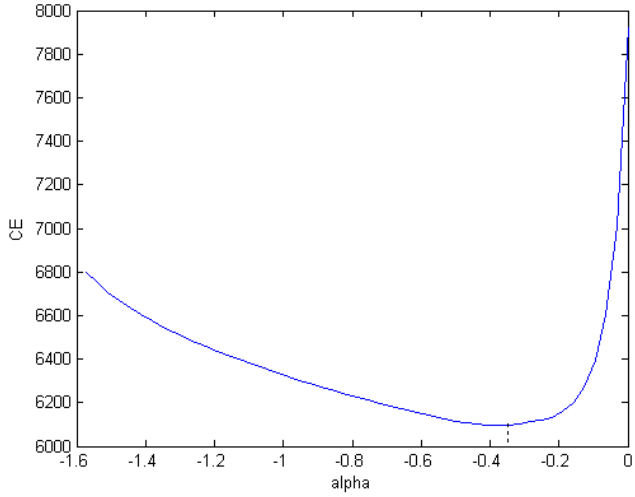


Figure 29. Cross-entropy error curve for Example 8.

Whereas the minimum entropy error occurs at $-\pi/2$:

$$\text{Error} = 0.3623$$

The minimum cross-entropy error occurs at -0.3456 and is:

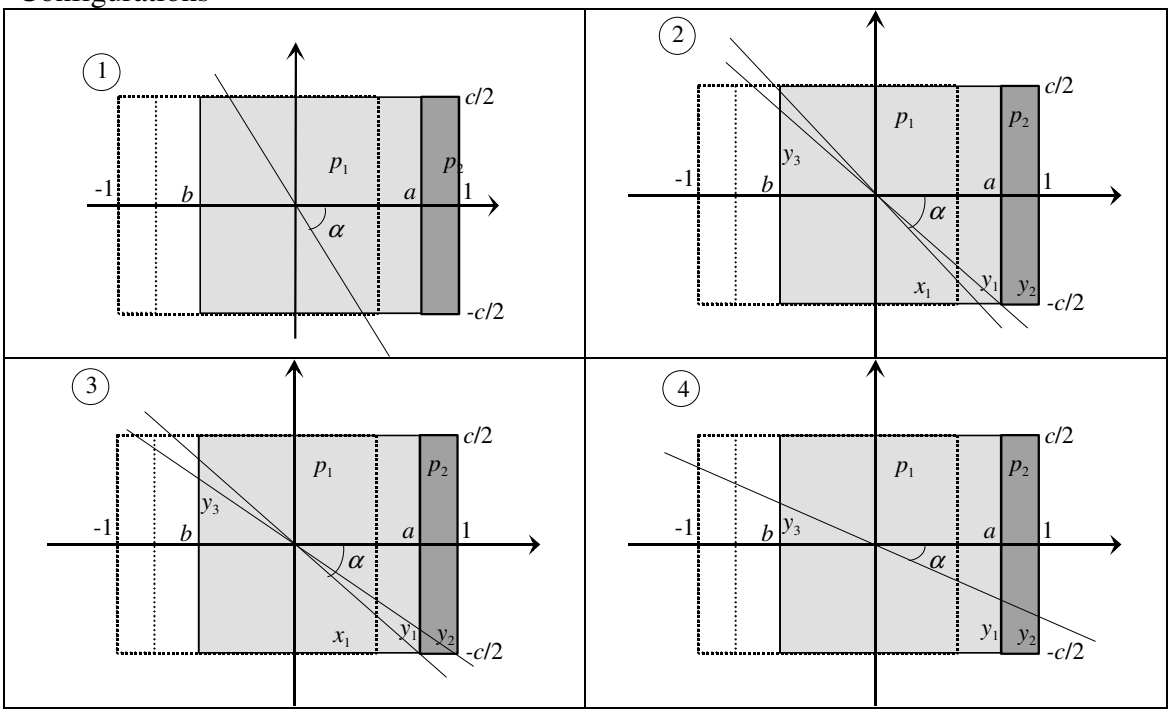
Error = 0.4105

10 Appendix A - Computation of the Error Curve

We assume the two rectangles with uniform distributions of either class having weights p_1 (large rectangle) and p_2 (small rectangle), with $p_1 + p_2 = 0.5$.

First configuration setting: $-b \le a$

Configurations



1) $\tan\alpha \le (c/2)/b$: $e = 2p_1 \dfrac{-b}{a-b}$

2) $(c/2)/b < \tan\alpha \le -(c/2)/a$: $x_1 = -b - \dfrac{c}{2}\dfrac{1}{\tan\alpha}$; $y_3 = \dfrac{c}{2} + b\tan\alpha$

$$e = 2p_1 \frac{x_1 y_3}{2} \frac{1}{c(a-b)}$$

3) $-(c/2)/a < \tan\alpha \le -(c/2)$: $y_1 = \dfrac{c}{2} + a\tan\alpha$; $x_1 = -y_1 \dfrac{1}{\tan\alpha}$; $y_3 = \dfrac{c}{2} + b\tan\alpha$

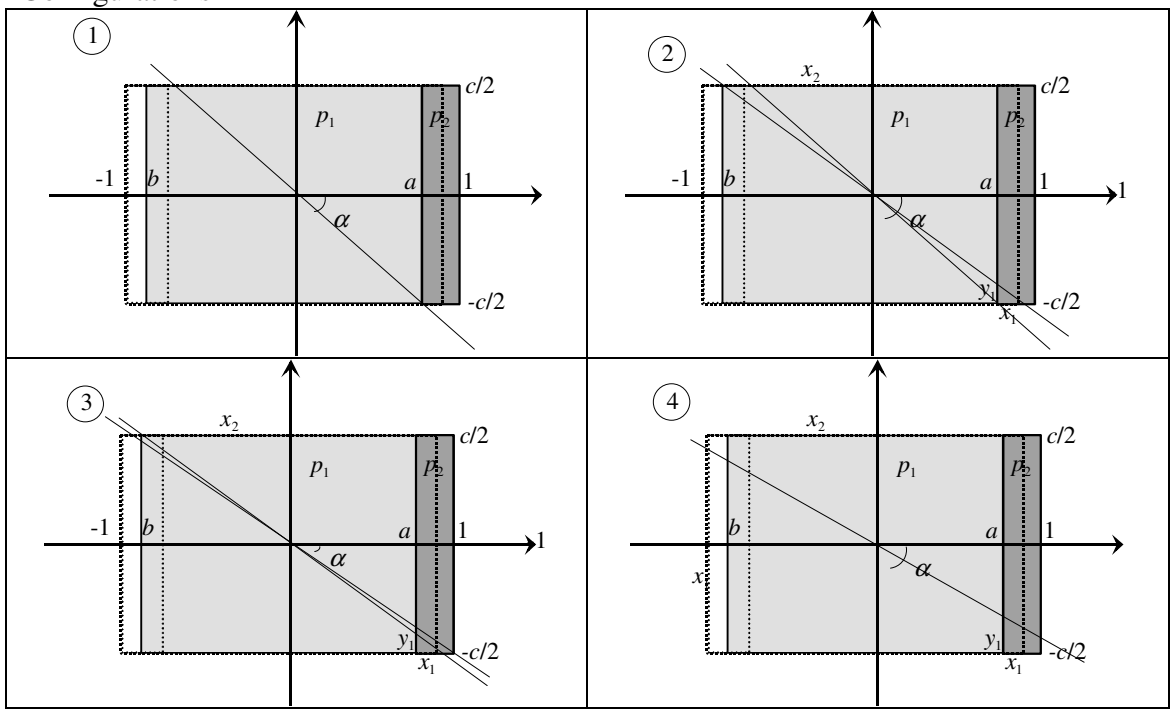
$$e = 2p_2 \frac{y_1 x_1}{2} \frac{1}{c(1-a)} + 2p_1 \frac{y_1 + y_3}{2} \frac{1}{c}$$

4) $\tan\alpha > -(c/2)$: $y_1 = \frac{c}{2} + a\tan\alpha$; $y_2 = \frac{c}{2} + \tan\alpha$; $y_3 = \frac{c}{2} + b\tan\alpha$

$$e = 2p_2 \frac{y_1 + y_2}{2}\frac{1}{c} + 2p_1 \frac{y_1 + y_3}{2}\frac{1}{c}$$

Second configuration setting: $a < -b \leq 1$

Configurations



1) $\tan\alpha \leq -(c/2)a$: $e = 2p_1 \frac{-b}{a-b}$

2) $-(c/2)a < \tan\alpha \leq (c/2)b$: $y_1 = \frac{c}{2} + a\tan\alpha$; $x_1 = -y_1 \frac{1}{\tan\alpha}$; $x_2 = 2a + x_1$

$$e = 2p_2 \frac{y_1 x_1}{2}\frac{1}{c(1-a)} + 2p_1\left(\frac{y_1 + c}{2}\frac{x_2}{c(a-b)} + \frac{-b-a-x_1}{a-b}\right)$$

3) $(c/2)b < \tan\alpha \leq -(c/2)$: $y_1 = \frac{c}{2} + a\tan\alpha$; $x_1 = -y_1 \frac{1}{\tan\alpha}$; $y_3 = \frac{c}{2} + b\tan\alpha$

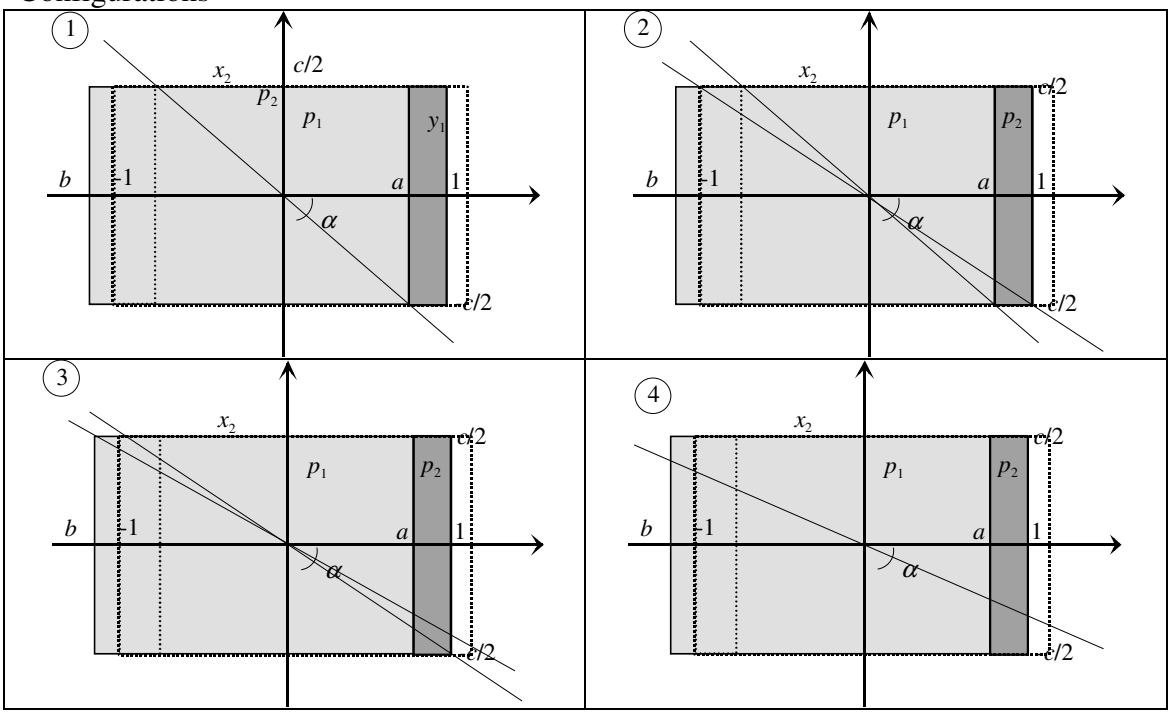
$$e = 2p_2 \frac{y_1 x_1}{2}\frac{1}{c(1-a)} + 2p_1 \frac{y_1 + y_3}{2}\frac{1}{c}$$

4) $\tan\alpha > -(c/2)$: $y_1 = \frac{c}{2} + a\tan\alpha$; $y_2 = \frac{c}{2} + \tan\alpha$; $y_3 = \frac{c}{2} + b\tan\alpha$

$$e = 2p_2 \frac{y_1 + y_2}{2}\frac{1}{c} + 2p_1 \frac{y_1 + y_3}{2}\frac{1}{c}$$

Third configuration setting: $-b > 1$

Configurations



1) $\tan\alpha \le -(c/2)a$: $e = 2p_1\dfrac{-b}{a-b}$

2) $-(c/2)a < \tan\alpha \le -(c/2)$: $y_1 = \dfrac{c}{2} + a\tan\alpha$; $x_1 = -y_1\dfrac{1}{\tan\alpha}$; $x_2 = 2a + x_1$

$$e = 2p_2\frac{y_1 x_1}{2}\frac{1}{c(1-a)} + 2p_1\left(\frac{y_1 + c}{2}\frac{x_2}{c(a-b)} + \frac{-b-a-x_1}{a-b}\right)$$

3) $-(c/2) < \tan\alpha \le (c/2)b$: $y_1 = \dfrac{c}{2} + a\tan\alpha$; $y_2 = \dfrac{c}{2} + \tan\alpha$; $x_1 = a - \dfrac{c}{2}\dfrac{1}{\tan\alpha}$

$$e = 2p_2\frac{y_1 + y_2}{2}\frac{1}{c} + 2p_1\left(\frac{y_1 + c}{2}\frac{x_1}{c(a-b)} + \frac{-b + \dfrac{c}{2}\dfrac{1}{\tan\alpha}}{a-b}\right)$$

4) $\tan\alpha > (c/2)b$: $y_1 = \dfrac{c}{2} + a\tan\alpha$; $y_2 = \dfrac{c}{2} + \tan\alpha$; $y_3 = \dfrac{c}{2} + b\tan\alpha$

$$e = 2p_2\frac{y_1 + y_2}{2}\frac{1}{c} + 2p_1\left(\frac{y_1 + y_3}{2}\right)\frac{1}{c}$$

11 Appendix B - 0ln0

We want to compute $\lim_{x\to 0} x\ln x$. But $x\ln x = \frac{\ln x}{1/x}$. Therefore, applying L'Hôpital's rule:

$$\lim_{x\to 0} x\ln x = \lim_{x\to 0} \frac{1/x}{-1/x^2} = 0$$