ON THE DIVERGENCE BETWEEN TWO DISTRIBUTIONS AND THE
PROBABILITY OF MISCLASSIFICATION OF SEVERAL DECISION RULES

Godfried T. Toussaint
School of Computer Science
McGill University
Montreal, Quebec
Canada

Abstract
Sharp lower bounds are derived for the
divergence between two distributions and the probabil-
ities of misclassification of three decision rules.
The three decision rules considered are the optimal
Bayes rule, the nearest neighbour rule, and the "pro-
portional prediction" randomized decision rule. It is
shown that the randomized rule yields a probability of
misclassification equal to the asymptotic nearest
neighbour error rate. The bound between the Bayes
error rate and the divergence is more general than the
Kullback bound and, unlike the latter, is distribution
free. The bounds are used to obtain sharp inequalities
between measures of probabilistic dependence between
features and classes in the multi-class pattern recog-
nition problem. The bounds lead to sharp inequalities
between the divergence and various information and
distance measures found in the literature. Finally,
the divergence is related to the least-mean-square-
error design criterion in pattern recognition.

1. Introduction

Consider the two-category pattern class-
ification problem. Let \( P(X/C_1) \) denote the class-
conditional probability density function of the feature
vector \( X \) conditioned on the class \( C_i \), \( i=1,2 \). The
Bhattacharyya coefficient and the divergence are de-
finite, respectively, by

\[
\rho = \int \sqrt{P(X/C_1) P(X/C_2)} \, dX
\]

and

\[
J = \int \left[ P(X/C_1) - P(X/C_2) \right] \log\left( \frac{P(X/C_1)}{P(X/C_2)} \right) \, dX.
\]

These measures are well known in the pattern recog-
nition literature [1] and are useful for feature selec-
tion when the underlying distributions are Gaussian
because they are much easier to evaluate than the error
probability.

Let class \( C_i \) occur with a priori probability
\( \pi_i, i=1,2 \), \( \pi_1 + \pi_2 = 1 \). It is useful to define more
general measures than (1) and (2) above, as follows:

\[
P(\pi_1, \pi_2) = \sqrt{\pi_1 \pi_2} \int \sqrt{P(X/C_1) P(X/C_2)} \, dX
\]

\[
= \int P(X) \sqrt{P(C_1/X) P(C_2/X)} \, dX
\]

(3)

and

\[
J(\pi_1, \pi_2) = \int \left[ \pi_1 P(X/C_1) - \pi_2 P(X/C_2) \right] \log\left( \frac{\pi_1 P(X/C_1)}{\pi_2 P(X/C_2)} \right) \, dX
\]

\[
= \int P(X) \left[ P(C_1/X) - P(C_2/X) \right] \log\left( \frac{P(C_1/X)}{P(C_2/X)} \right) \, dX.
\]

(4)

It follows that \( p(1/2,1/2) = \rho/2 \) and \( J(1/2,1/2) = J/2 \). In
(3) and (4) \( P(X) \) is the mixture distribution and is given by

\[
P(X) = \pi_1 P(X/C_1) + \pi_2 P(X/C_2).
\]

In this paper, the divergence is related to the proba-
bilities of misclassification of three well known deci-
sion rules. These relationships are important when
one would like to know what performance can be expect-
ed from a decision rule when features have been selec-
ted using the divergence. The first decision rule is
the optimal Bayes rule. Given a feature vector \( X \) from
some unknown pattern \( P \), \( P \) is classified as belonging
to class \( C_i \) if \( P(C_i/X) > P(C_j/X), i \neq j \). This
rule gives the minimum possible probability of mis-
classification [2] which is given by

\[
P_e = \int \min \left( P(X/C_i), \pi_1 \right) \, dX, \quad i=1,2
\]

\[
= \int P(X) \min \left( P(C_1/X), P(C_2/X) \right) \, dX
\]

\[
= \int P(X) P_e(X) \, dX.
\]

(5)

The second decision rule considered here is the near-
pest neighbour rule (NN-rule). Let \( (X_0, \theta_0) = \)
\( \{X_1, \theta_1; X_2, \theta_2; \ldots; X_N, \theta_N\} \) be the set of \( N \) pattern
samples available, where \( X_i \) and \( \theta_i \) denote, respectiv-
ely, the feature vector or measurement information and
the label or classification information of the ith
pattern sample. It is assumed that each \( \theta_i \) associated
with \( X_i \) is the correct label, i.e., the pattern
samples have been correctly pre-classified. Let
\( (X_n, \theta_n) \in (X_i, \theta) \) to be the sample nearest to the un-
known \( X \). \( P \) is then classified as belonging to the
The third decision rule under investigation is in the randomized decision rule. Let the class-conditional distributions be known as in the deterministic Bayes rule. Given a feature vector $X$ from some unknown pattern $P$, $P$ is classified as belonging to Class $C_i$, $i=1,2$, by a flip of a biased coin which indicates $C_i$ with probability $P(C_i|X)$. This type of decision rule tends to produce a distribution of class assignments more similar to the original distribution than does the deterministic Bayes rule and is also known as proportional prediction [4]. The probability of misclassification using this rule, denoted by $R$ for reasons that will become apparent, can be derived as follows.

For any given value of $X$, $C_1$ occurs with probability $P(C_1|X)$ and it is decided to belong to class $C_2$ with probability $1-P(C_1|X)$. Similarly, $C_2$ occurs with probability $P(C_2|X)$ and it is decided to belong to class $C_1$ with probability $1-P(C_2|X)$. Hence, for a given value of $X$ the resulting probability of misclassification is given by

$$R(X) = P(C_2|X)[1-P(C_2|X)] + P(C_1|X)[1-P(C_1|X)] = 2P(C_1|X)P(C_2|X)$$

Taking the expected value of (7) with respect to $P(X)$ yields

$$R = \int P(X) R(X) \, dX,$$

which is the same as the asymptotic nearest neighbor error rate of (6). This equivalence between the NN-rule and the proportional prediction randomized rule (PFR-rule) has not been noted in the literature and provides added insight into the deterministic NN-rule. When an unknown $X$ is far from the decision boundary into the region for Class $C_1$ the NN-rule will almost always choose class $C_1$ unless a maverick is near $X$. In the PFR-rule mavericks are explained by the fact that $P(C_1|X)$ is hardly ever equal to one. On the other hand, when $X$ lies around the decision boundary it is likely to have nearest neighbors of either class. In terms of the PFR-rule one chooses $C_1$ with probability $P(C_1|X)$ which is close to 0.5 when $X$ is close to the decision boundary.

In this paper a generalized version of the inequality of Hoeffding and Wolfowitz [5] is derived. Using this inequality lower bounds are derived for $P_e$ and $R$ in terms of $J$. It is shown that the bounds for $P_e$ are tighter than existing bounds. In addition, sharper lower bounds are derived for the divergence $J$ in terms of $P_e$ and $R$. Their relation to the Kullback-Leibler bound is discussed. The bounds are applied to dependence measures between features and classes, equivocation measures, and distance measures found in the literature. The divergence is finally also related to the least-mean-square-error design criterion.

2. An Inequality Between $J(\pi_1, \pi_2)$ and $\rho(\pi_1, \pi_2)$

The divergence between two distributions occurring with prior probabilities $\pi_1$ and $\pi_2$ can be written as

$$J(\pi_1, \pi_2) = -\pi_1 E_1 \left\{ \log \frac{P(X|C_1)}{P(X|C_2)} \right\}$$

$$-\pi_2 E_2 \left\{ \log \frac{P(X|C_1)}{P(X|C_2)} \right\}$$

(9)

where $E_i$ denotes expected value with respect to $P(X|C_i)$. Since $\log x$ is a convex upward function ($\triangleleft$), Jensen's inequality applied to (9) gives

$$J(\pi_1, \pi_2) \geq -2 \pi_1 \log E_1 \left\{ \frac{P(X|C_1)}{P(X|C_2)} \right\}$$

$$-2 \pi_2 \log E_2 \left\{ \frac{P(X|C_1)}{P(X|C_2)} \right\},$$

which, in turn, yields

$$J(\pi_1, \pi_2) \geq -2 \pi_1 \log \left[ \rho(\pi_1, \pi_2)/\pi_1 \right]$$

$$-2 \pi_2 \log \left[ \rho(\pi_1, \pi_2)/\pi_2 \right].$$

(10)

Expanding (10) and recombining terms yields the desired result given by

$$J(\pi_1, \pi_2) \geq -2 \left[ H(\pi) + \log \rho(\pi_1, \pi_2) \right],$$

where $H(\pi)$ is the entropy function given by

$$H(\pi) = -\pi_1 \log \pi_1 - \pi_2 \log \pi_2.$$  

(11)

When $\pi_1 = \pi_2 = 1/2$, $H(\pi) = \log 2$ and (11) reduces to

$$J \geq -4 \log \rho$$

(13)

which is a well known inequality due to Hoeffding and Wolfowitz [5]. Hence (11) is a generalization of (13) to take into account the a priori probabilities.

3. Lower Bounds for $R$ and $J$

Although there exists a lower bound on $R(X)$ in terms of $J(X)$ no bounds are available in the literature, between $R$ and $J$. Horibe [6] showed that

$$R \geq 2 \left[ \rho(\pi_1, \pi_2) \right]^2,$$

from which it follows that

$$\log \rho(\pi_1, \pi_2) \leq \log \sqrt{R/2}.$$  

(14)

Substituting (14) into (11) yields
number of papers on feature selection and texts on pattern recognition [9]. It is given by
\[ P_e \geq (1/4) \exp \left\{ -J/2 \right\} \]
where the equality holds of \( J = -\infty \). This bound is illustrated in Fig. 2 where it is seen that it is a loose bound.

Cover and Hart [3] have shown that
\[ R \leq 2P_e (1-P_e) \]
and, hence, that
\[ R \leq 2P_e \]
Substituting (24) into (15) yields
\[ P_e \geq \exp \left\{ -2 H(\pi) - J(\pi_1, \pi_2) \right\} . \)
For \( \pi_1 = \pi_2 = 1/2 \), (25) reduces to (22) and, hence, (25) can be considered as a generalization of Kailath’s bound. Similarly, substituting (24) into (18) yields
\[ P_e \geq (1/4) \left\{ 1 - J(\pi_1, \pi_2) \right\} \]
For \( \pi_1 = \pi_2 = 1/2 \), (26) reduces to
\[ P_e \geq 1/4 - J/16 \]
which is illustrated in Fig. 2. For \( J \geq 4 \) it is useless, but for \( J \leq 3.2 \) it is sharper than Kailath’s bound and, hence, improves the latter when used in conjunction with it.

Tighter bounds than (25) and (26) can be obtained by using (23) rather than (24). Substituting
\[ (23) \]
into (15) yields
\[ P_e \geq (1/2) - \left( 1/2 \right) \sqrt{1-4 \exp\left\{-2 H(\pi) - J(\pi_1, \pi_2) \right\}} \]
Solving (28) for \( P_e \) yields
\[ P_e \geq (1/2) - \left( 1/2 \right) \sqrt{J(\pi_1, \pi_2)} / 8 \]
For \( \pi_1 = \pi_2 = 1/2 \), (29) reduces to
\[ P_e \geq (1/2) - \left( 1/2 \right) \sqrt{J} \]
where the equality holds for both \( J=0 \) and \( J=-\infty \). Similarly, substituting (23) into (18) and solving for \( P_e \) yields
\[ P_e \geq (1/2) - \left( 1/4 \right) \sqrt{J} \]
which for equal a priori probabilities reduces to
\[ P_e \geq (1/2) - \left( 1/4 \right) \sqrt{J} \]
where the equality holds when \( J=0 \). Bounds (30) and (32) are also illustrated in Fig. 2 which shows that (32) is sharper than (30) for \( J < 3.2 \).

A lower bound on \( J \) in terms of \( P_e \) which is sharper than all the above bounds can be derived as follows. From (4) it follows that
\[ J(\pi) = [ P(C_1/X) - P(C_2/X) ] \log \left\{ \frac{P(C_1/X)}{P(C_2/X)} \right\} \]

4. Lower Bounds for \( P_e \) and \( J \)

There exists in the literature a lower bound on \( P_e \) in terms of \( J \). It appears to have been derived first by Kailath [8] and has since appeared in a
Also, from (5) it follows that
\[ P_e(X) = \min \{ P(C_1/X), P(C_2/X) \} \]  
(34)
The crucial step in the derivation of the bound is the realization that, since \( J(X) \) is symmetrical with respect to \( P(C_1/X) \) and \( P(C_2/X) \), it can be expressed in terms of \( P_e(X) \) as
\[ J(X) = \left\{ 2 \ P_e(X) - 1 \right\} \log \left\{ \frac{P_e(X)}{1 - P_e(X)} \right\} \]  
(35)
Now consider the function
\[ f(x) = (2x - 1) \log \left( \frac{x}{1-x} \right) \]
in the interval \( 0 \leq x \leq 1/2 \). The first derivative of \( f(x) \) with respect to \( x \) is given by
\[ df(x) = \frac{2x - 1 + 2x - 1 + 2 \log \left( \frac{x}{1-x} \right)}{x(1-x)} \]  
(36)
The second derivative is given by
\[ \frac{d^2f(x)}{dx^2} = \frac{1-2x}{x^2} + \frac{4}{x} - \frac{4}{x} - \frac{1-2x}{(1-x)^2} \]  
(37)
It can easily be shown that (37) is non-negative.
Hence \( f(x) \) and (35) are convex downward functions. For a convex downward function Jensen's inequality is given by
\[ E \{ f(X) \} \geq f(E \{ X \}) \]  
(38)
where \( E \) denotes expected value. Taking the expected value of both sides of (35) with respect to \( P(X) \) and using (38) yields the desired bound given by
\[ J(\tau_1, \tau_2) \geq (2 \ P_e - 1) \log \left\{ P_e/(1-P_e) \right\} \]  
(39)
For equal prior probabilities (39) reduces to
\[ J \geq 2(2 \ P_e - 1) \log \left\{ P_e/(1-P_e) \right\} \]  
(40)
where the equality holds for both \( P_e=0 \) and \( P_e=1/2 \). As illustrated in Fig. 2, (40) is the sharpest inequality between \( J \) and \( P_e \) but has the disadvantage that it cannot be solved for \( P_e \) as a function of \( J \). For a proof that (40) is sharper than (30) and (32) see Appendix B.

From (6) it follows that
\[ R(X) = 2 \ P(C_1/X) \ P(C_2/X) \]  
(41)
which can be written as
\[ R(X) = 2 \ P_e(X) \left\{ 1 - P_e(X) \right\} \]  
(42)
where \( P_e(X) \) is given by (34). Solving the \( P_e(X) \) yields
\[ P_e(X) = (1/2) - \sqrt{(1/4) - R(X)/2} \]  
(43)
It can easily be shown that (43) is a convex downward function of \( R(X) \). Taking expected values, with respect to \( P(X) \), of both sides of (43) yields, using Jensen's inequality,
\[ P_e \geq \left( \frac{1}{2} \right) - \sqrt{\left( \frac{1}{4} \right) - R/2} \]  
(44)
Substituting (44) into (39) yields (20).

5. Relation to Kullback Bounds

The Kullback-Liebler numbers [10] are given by
\[ I(1,2) = \int P(X/C_1) \log \frac{P(X/C_1)}{P(X/C_2)} \, dx \]  
(45)
and
\[ I(2,1) = \int P(X/C_2) \log \frac{P(X/C_2)}{P(X/C_1)} \, dx \]  
(46)
Let \( P_{e1} \) denote the probability of misclassification given class \( C_1 \), \( 1=1,2 \), where \( P_e = \tau_1 P_{e1} + \tau_2 P_{e2} \).
The Kullback bounds are given by [8], [10], [11],
\[ I(1,2) \geq P_{e1} \log \frac{P_{e1}}{(1-P_{e2})} + (1-P_{e1}) \log \frac{(1-P_{e1})}{P_{e2}} \]  
(47)
and
\[ I(2,1) \geq P_{e2} \log \frac{P_{e2}}{(1-P_{e1})} + (1-P_{e2}) \log \frac{(1-P_{e2})}{P_{e1}} \]  
(48)
When the distributions are such that
\[ f_{\gamma_{X/C_1}} P(X/C_1) \, dx = f_{\gamma_{X/C_2}} P(X/C_2) \, dx \]  
(49)
where \( \gamma_{X/C_1} = \{ \gamma_X : P(X/C_1) > P(X/C_2) \} \),
\( i=1,2, i \neq j \), and \( \gamma_X \)
is the entire feature space, then \( P_{e1} = P_{e2} = P_e \). For example, (49) is true for Gaussian distributions with equal covariance matrices. Adding (47) and (48), substituting \( P_{e1} = P_{e2} = P_e \) and using the fact that \( I(1,2) + I(2,1) = J \) yields
\[ J \geq 2(2P_e - 1) \log \frac{P_e/(1-P_e)}{1-P_e} \]  
(50)
which is a special case of (39). It is nice to know that assumption (49) is not needed and that (50) actually holds in general.

6. Application to Dependence Measures

Consider the \( M \)-class problem. A measure of the dependence between features and classes can be obtained by measuring the distance, in some sense, between the joint probability distribution \( P(X,C) \) and the product of the marginals \( P(X)P(C) \). Vilmsen [12] considers various measures of probabilistic dependence in this way and relates them to the probability of misclassification \( P_e \). Two measures considered in [12] are the Kolmogorov dependence, first proposed by Hoeffding [13] and, given by
\[ D_K(X,C) = \frac{M}{i \neq j} \int \left| P(X,C_i) - P(X) \tau_i \right| \, dx \]  
(51)
and the Joshi dependence, first proposed as a measure of channel capacity, Joshi [14], and, given by
\[ D_J(X,C) = \frac{M}{i \neq j} \int \left( P(X,C_i) - P(X) \tau_i \right) \log \frac{P(X,C_i)}{P(X) \tau_i} \, dx \]  
(52)
The bounds between \( J \) and \( P_e \), for the 2-class problem derived in section 4, can be used to form sharp inequalities between the above dependence measures for
the M-class problem.

It can easily be shown [15] that, for equal a priori probabilities,

\[ P_e = (1/2) - V/4, \]

where \( V \) is the Kolmogorov variational distance given by

\[ V = \int_{-\infty}^{\infty} \left| P(X/C_1) - P(X/C_2) \right| \, dx. \]

Substituting (53) into (31), (32) and (40) yields, respectively,

\[ V \leq 2 \left[ 1 - \exp(-J/2) \right]^{1/2}, \]

\[ V \leq \left( J/2 \right)^{1/2}, \]

and

\[ J \geq V \log \left( \frac{2 + V}{2 - V} \right). \]

Realizing that \( D_K(X,C) \) are distance measures between two distributions in a continuous-discrete space of dimensionality one greater than the dimensionality of \( X \), allows one to write (55)-(57), respectively, in the following way.

\[ D_J(X,C) \geq D_K(X,C) \log \left[ \frac{2 + D_K(X,C)}{2 - D_K(X,C)} \right], \]

\[ D_K(X,C) \leq 2 \left[ 1 - \exp \left[ -D_J(X,C)/2 \right] \right]^{1/2}, \]

\[ D_K(X,C) \leq \left[ D_J(X,C) \right]^{1/2}, \]

The Kolmogorov dependence \( D_K(X,C) \) can also be related to the expected divergence \( J \) which is given by

\[ J = \sum_{i=1}^{M} \sum_{j=1}^{N} P_{i,j} \log \left( \frac{P_{i,j}}{P_{i,C_j}} \right), \]

where \( P_{i,j} \) is the divergence between \( P(X/C_1) \) and \( P(X/C_2) \).

It was shown in [16] that

\[ J = 2 D_J(X,C). \]

This relation supports Vilmansen's conjecture [17] that there is a close relationship between the dependence of features and classes and the distance between class-conditional distributions. Substituting (62) into (58), (39) and (60) yields, respectively,

\[ J \geq 2 D_K(X,C) \log \left[ \frac{2 + D_K(X,C)}{2 - D_K(X,C)} \right], \]

\[ D_K(X,C) \leq 2 \left[ 1 - \exp(-J/4) \right]^{1/2}, \]

and

\[ D_K(X,C) \leq \left( J/2 \right)^{1/2}, \]

where the equality holds when classes and features are independent.

One measure of dependence not considered in [12] can be developed from the asymptotic nearest neighbour error rate \( R \). For equal a priori probabilities \( R \) is given by

\[ R = \int_{-\infty}^{\infty} \frac{P(X/C_1) P(X/C_2)}{P(X/C_1) + P(X/C_2)} \, dx, \]

which, in a sense, measures the distance between \( P(X/C_1) \) and \( P(X/C_2) \). Hence, a new measure of dependence can be defined as

\[ D_R(X,C) = \int_{-\infty}^{\infty} \frac{P(X,C_j) P(X)}{P(X/C_1) + P(X/C_2)} \, dx. \]

Furthermore, from the fact that [3]

\[ P_e \leq R \leq 2 P_e (1 - P_e), \]

using (53) and similar arguments as above, it follows that

\[(1/2) - D_K(X,C)/4 \leq D_R(X,C) \leq (1/2) - (1/8) \left( \frac{1}{D_K(X,C)} \right)^2, \]

where the equalities hold for \( D_K(X,C) = 0 \), i.e., when the features and classes are independent.

7. Application to Equivocation Measures

Shannon's measure of equivocation is the most well known and, for the 2-class problem, is given by

\[ H(C/X) = - \int_{-\infty}^{\infty} P(C_1/X) \log P(C_1/X) \, dx. \]

Not as well known is Vajda's quadratic equivocation [18] given by

\[ Q(C/X) = - \int_{-\infty}^{\infty} P(C_1/X) P(C_1/X) \, dx. \]

Recently, Toussaint [16], [19], [20] proposed a family of equivocation measures given by

\[ M_k(C/X) = \int_{-\infty}^{\infty} P(C_1/X) \log \left( \frac{P(C_1/X)}{k} \right) \, dx, \]

where \( k = 2(k+1)/(2k+1) \) and \( k = 0, 1, 2, \ldots \). Of particular interest here is \( M_0(C/X) \) given by

\[ M_0(C/X) = \int_{-\infty}^{\infty} P(C_1/X) \log \left( \frac{P(C_1/X)}{1 - P(C/X)} \right) \, dx. \]

It was shown in [16] that \( M_0(C/X) \) is related to the asymptotic nearest neighbour error rate by the relation

\[ R = (1/2) - M_0(C/X). \]

It also follows that

\[ M_0(C/X) = 1 - Q(C/X). \]

Hence, the information measure \( Q(C/X) \), which is obtained by approximating \( \log x \) by \( (x-1) \) in Shannon's logarithmic equivocation, is also a distance measure (the harmonic mean between \( P(X/C_1) \) and \( P(X/C_2) \) ) as well as the asymptotic nearest neighbour error rate, and the probability of error of the proportional prediction randomized decision rule. Since \( \log x \leq x-1 \), it follows that \( R \) is bounded above by Shannon's equivocation, i.e.,

\[ R \leq H(C/X). \]

Substituting (74) and (75) into the bounds on \( R \) in section 3 gives sharp inequalities between the divergence \( J \) and the various equivocation measures. For example, substituting (75) into (15), (18), and (20) yields, respectively,
\[
Q(C/X) \geq 2 \exp[-2 H(c) - J(\tau_1, \tau_2)] ,
\]
\[
Q(C/X) \geq \frac{1}{2} \left[ 1 - J(\tau_1, \tau_2)/2 \right] ,
\]
and
\[
J(\tau_1, \tau_2) \geq \sqrt{1 - 2 Q(C/X)} \log \left[ \frac{1 + \sqrt{1 - 2 Q(C/X)}}{1 - \sqrt{1 - 2 Q(C/X)}} \right].
\]

8. Application to Distance Measures

Ito [21] proposed a family of distance measures, called the Q-function, given by

\[
Q_n = (1/2) - (1/2) \int P(X) \left[ P(C_1/X) - P(C_2/X) \right]^n \; dX ,
\]
where \( n = 1/(2n+1) \) and \( n \) is a natural number. Of particular interest is \( Q_0 \) given by

\[
Q_0 = (1/2) - d/2
\]
where

\[
d = \int P(X) \left[ P(C_1/X) - P(C_2/X) \right]^2 \; dX .
\]

Ito [21] showed that

\[
Q_0 = R ,
\]
\[
P_e = P_e ,
\]
and

\[
Q_{n+1} = Q_n .
\]

Substituting these results into the lower bounds for \( R \) relates the \( Q \)-function to the divergence.

Lissack and Fu [22] have investigated feature selection and estimation of misclassification using the separability measure

\[
J_a = \int P(X) \left[ P(C_1/X) - P(C_2/X) \right]^a \; dX \quad (85)
\]
for \( a > 0 \). It can easily be shown that

\[
P_e = (1/2) - J_1/2 ,
\]
and

\[
R = (1/2) - J_2/2 .
\]

Hence, substituting (86) and (87) into the results of sections 3 and 4 relates \( J_a \) to the divergence \( J \).

Deviyer [23], [24] has recently done a lot of work on the so-called Bayesian distance given by

\[
B(C/X) = \int P(X) \left[ P(C_1/X) \right]^2 \; dX .
\]

It is obvious that

\[
B(C/X) = 1 - R .
\]

Hence, using the results of section 3 yields sharp inequalities between the Bayesian distance and the divergence. For example, letting \( B \) denote \( B(C/X) \), to simplify notation, and substituting (89) into (21) yields

\[
J \geq 2 \sqrt{2B - 1} \log \left[ \frac{1 + \sqrt{2B - 1}}{1 - \sqrt{2B - 1}} \right] .
\]

9. Concluding Remarks

It has been shown that the probability of misclassification of the proportional-prediction randomized decision rule is equivalent to the error rate of the deterministic nearest neighbour rule, asymptotically. Previously, no bounds were available for \( R \) and the divergence \( J \). In this paper better lower bounds are given for \( R \) and \( J \). The tightest bound is given by (21). However, for feature evaluation using \( J \), (16) and (19) are more useful. Letting

\[
R_1 = (1/2) \exp[-J/2],
\]
and

\[
R_2 = (1/2) - J/8 ,
\]
the best lower bound recommended for future use is

\[
R \geq \max \left\{ R_1, R_2 \right\} .
\]

Similar comments hold true for \( P_e \). For Gaussian distributions, an upper bound on \( P_e \) in terms of \( J \) is available and is given by [25]

\[
P_e \leq (1/2) (3/4)^{-1/4} .
\]

Letting

\[
L_1 = (1/2) - (1/2) \sqrt{1 - \exp(-J/2)} ,
\]
and

\[
L_2 = (1/2) - (1/4) \sqrt{J} ,
\]
from (30) and (32), the best available lower bound to complement (91) above is given by

\[
P_e \geq \max \left\{ L_1, L_2 \right\} ,
\]
which is greatly superior to the previous available bound, (22).

A final comment is in order as regards the well known least-mean-square-error (LMSE) design criterion [26] which has received a great deal of attention in the pattern recognition literature. Deviyer [27] has shown that for a certain class of risk functions the LMSE criterion is equal to \( R \). Under these conditions, (21) shows that minimizing the LMSE is equivalent to maximizing a lower bound on the divergence \( J \).

Appendix A

Equation (19) can be written in the form

\[
J \geq 4(1 - 2R) .
\]

To show that (21) is sharper than (19) it must be proved that

\[
2 \sqrt{1 - 2R} \log \left[ \frac{1 + \sqrt{1 - 2R}}{1 - \sqrt{1 - 2R}} \right] \geq 4(1 - 2R) .
\]

Making use of the transformation \( [1 - 2R]^{1/2} = x \), \( 0 \leq x \leq 1 \), it must follow that

\[
2x \log \left[ (1 + x)/(1 - x) \right] \geq 4x^2 ,
\]
which in turn yields

\[
\log \left[ (1 + x)/(1 - x) \right] \geq 2x .
\]

It is known that

\[
\log \left[ (1 + x)/(1 - x) \right] = 2 \sum_{k=1}^{\infty} \left[ 1/(2k - 1) \right] x^{2k-1} .
\]
for $x^2 > 1$. Since $x > 0$, all terms in (A4) are nonnegative and it follows that for $k > 1$ (A4) reduces to (A3), proving the result.

Equation (16) can be written in the form

$$J \geq 2 \log(2k) - 2 \log(2) \tag{A5}$$

To show that (21) is sharper than (16) it must hold true that

$$2 \log \left( \frac{1-x}{1-x} \right) \geq 2 \log \left( 1 - x^2 \right) \tag{A6}$$

where $x$ is as above. Making use of the transformation

$$x \rightarrow 1/\sqrt{x}, \quad 1/2 \leq y \leq 1,$$

which must hold true that

$$2 \log \left( \frac{1-y}{1-y} \right) \geq 2 \log \left( \frac{1+y}{1+y} \right) \tag{A7}$$

Expanding (A7) and recombining terms results in

$$H(y, 1-y) \leq \log 2 \tag{A8}$$

where $H(y, 1-y)$ is the entropy function. The maximum of $H(y, 1-y)$ occurs for $y = 1/2$ and is given by $\log 2$, thus proving the desired result.

Appendix B

Equation (30) can be written in the form

$$J \geq 2 \log \left( \frac{4 \alpha}{1 - \alpha} \right) \tag{B1}$$

To prove that (40) is sharper than (30) it must be shown that

$$2 \log \left( \frac{P_e}{1-P_e} \right) \geq 2 \log \left( \frac{P_e}{1-P_e} \right) \tag{B2}$$

which is of the same form as (A7), thus proving the result.

Equation (32) can be written in the form

$$J \geq 4 \left( 1 - 2 P_e \right)^2 \tag{B3}$$

To prove that (40) is sharper than (32) it must be shown that

$$2 \log \left( \frac{1-x}{1-x} \right) \geq 2 \log \left( \frac{1-x}{1-x} \right) \tag{B4}$$

for $0 \leq x \leq 1/2$. Using the transformation

$$x = 1/(z+1), \quad 1 \leq z \leq e,$$

it must be shown that

$$\log z \geq 2 \left( \frac{e-1}{e+1} \right) \tag{B5}$$

It is known that for $z > 0$

$$\log z = \sum_{k=1}^{\infty} \frac{1}{k} \left( \frac{1}{2k-1} \right)^{2k-1} \tag{B6}$$

for $k \geq 1$, all terms in (B5) are non-negative. Hence, for $k \geq 1$ (B5) reduces to (B4), thus proving the result.

REFERENCES


---

**Figure 1**

**Figure 2**