A SAMPLE SIZE DEPENDENT ERROR BOUND

Robert P.W. Duin
Department of Physics, Delft University of Technology
Delft, The Netherlands

Abstract

An upperbound is derived for the classification error made by a Bayes discriminant function which is estimated by a finite learning set. This upperbound is expressed in the Bayes error made in the infinite sample case and the estimation errors for the distributions of the classes. The expectation of such an estimation error is a function of the number of learning samples. The upperbound for the expected error can therefore be written as a function of the sample size. This makes it possible to compute the number of learning samples that guarantees a certain accuracy in the discriminant function. For the cases of a general measurement space and of normal distributed classes these numbers are computed as a function of the measurement complexity and the dimensionality respectively.

Introduction

In the pattern recognition literature error bounds are intensively studied [1]. The main purpose of these studies is the construction of easily computable error bounds in the case of known class distributions. These error bounds are therefore sample size independent. They don't take into account the error made by the estimation of the distribution. For answering questions like: What is the error caused by a finite learning set, or: What number of learning samples should be used in order to reach a certain accuracy, these error bounds are useless.

Effects of the sample size upon the accuracy of the discriminant function are previously studied. Cover[2] and Foley[3] give lower bounds of two and respectively about five times as many samples as features. Extremely large figures are given by Hughes[4] and Abend et al.[5]. They compute the optimal measurement complexity for a given sample size. The resulting sample size for which a given measurement complexity is optimal is very large due to the general approach in which nearly each distribution is allowed.

In this paper an upperbound for the expectation of the discriminant error is given which is expresses in the Bayes error for the case of infinite sample size and in the expected errors made in the estimation of the class distributions. The expected estimation error, and therefore the upperbound is a function of the sample size. This makes it possible to compute the maximum number of learning samples necessary for a given expected value of the discriminant error.

General formulation

Let $f_A(x)$ and $f_B(x)$ be the two density functions of the classes A and B; let $\hat{f}_A(x)$ and $\hat{f}_B(x)$ be their estimates. The Bayes error is given by

$$\varepsilon^\star = \tfrac{1}{2} \int_{f_A(x)\leq f_B(x)} f_A(x)dx + \tfrac{1}{2} \int_{f_A(x)>f_B(x)} f_B(x)dx \qquad (1)$$

We assume equal class probabilities. Expression (1) is equivalent with

$$\varepsilon^\star = \tfrac{1}{2} \int_{\forall x} \min\{f_A(x),\ f_B(x)\}dx \qquad (2)$$

The error made by the estimates $\hat{f}_A(x)$ and $\hat{f}_B(x)$ will be defined by

$$e_A = \tfrac{1}{2} \int_{\forall x} |\hat{f}_A(x) - f_A(x)|dx \qquad (3)$$

which is equivalent with

$$e_A = 1 - \int_{\forall x} \min\{\hat{f}_A(x),\ f_A(x)\}dx \qquad (4)$$

and by

$$e_B = \tfrac{1}{2} \int_{\forall x} |\hat{f}_B(x) - f_B(x)|dx \qquad (5)$$

which is equivalent with

$$e_B = 1 - \int_{\forall x} \min\{\hat{f}_B(x),\ f_B(x)\}dx \qquad (6)$$

The definitions are such that $0 \leq e_A \leq 1$ and $0 \leq e_B \leq 1$. For perfect estimates e_A and e_B are zero, for bad estimates they approach or are equal to one. The error made by the difference,

$$\hat{S}(x) = \hat{f}_A(x) - \hat{f}_B(x) \qquad (7)$$

is
$$\varepsilon = \tfrac{1}{2} \int_{\hat{S}<0} f_A(x)dx + \tfrac{1}{2} \int_{\hat{S}\geq 0} f_B(x)dx \qquad (8)$$

We can rewrite this, if

$$S(x) = f_A(x) - f_B(x) \qquad (9)$$

by
$$\varepsilon = \tfrac{1}{2} \int_{S<0} f_A(x)dx - \tfrac{1}{2} \int_{\substack{S<0 \\ \hat{S}\geq 0}} f_A(x)dx + \tfrac{1}{2} \int_{\substack{S\geq 0 \\ \hat{S}<0}} f_A(x)dx$$

$$+ \tfrac{1}{2} \int_{S\geq 0} f_B(x)dx - \tfrac{1}{2} \int_{\substack{S\geq 0 \\ \hat{S}<0}} f_B(x)dx + \tfrac{1}{2} \int_{\substack{S<0 \\ \hat{S}\geq 0}} f_B(x)dx \qquad (10)$$

$$= \tfrac{1}{2} \int_{S<0} f_A(x)dx + \tfrac{1}{2} \int_{S\geq 0} f_B(x)dx - \tfrac{1}{2} \int_{\substack{S<0 \\ \hat{S}\geq 0}} S(x)dx + \tfrac{1}{2} \int_{\substack{S\geq 0 \\ \hat{S}<0}} S(x)dx \qquad (11)$$

From (1) and (9) it follows that the first two terms are together $\varepsilon^\star$. If we define an area V in which the classes are non-optimally classified.

$$V = \{x: (S(x)<0 \cap \hat{S}(x)\geq 0) \cup (S(x)\geq 0 \cap \hat{S}(x)<0)\} \qquad (12)$$

then (11) can be written as

$$\varepsilon = \varepsilon^\star + \tfrac{1}{2} \int_{x\in V} |S(x)|dx \qquad (13)$$

$$= \varepsilon^\star + \tfrac{1}{2} \int_{x\in V} |f_A(x) - f_B(x)|dx \qquad (14)$$

For $x \in V$ the following inequality holds

$$|\hat{f}_A(x) - f_A(x)| + |\hat{f}_B(x) - f_B(x)| \geq |f_A(x) - f_B(x)| \qquad (15)$$

For the proof we distinguish the two cases

a) $S(x) < 0,\ \hat{S}(x) \geq 0$ (16)

so $-S(x) \leq \hat{S}(x) - S(x)$ (17)

Because both terms are positive, (17) is also true for the absolute values

$$|S(x)|=|f_A(x)-f_B(x)|\leq|\hat{S}(x)-S(x)|$$
$$\leq|\hat{f}_A(x)-\hat{f}_B(x)-f_A(x)+f_B(x)|$$
$$\leq|\hat{f}_A(x)-f_A(x)|+|\hat{f}_B(x)-f_B(x)| \qquad (18)$$

which proves (15)

b) $S(x) \geq 0,\ \hat{S}(x) < 0$ (19)

The proof in this case is similar to that in a. By using (15), (14) can be written as

$$\varepsilon\leq\varepsilon^{\star}+\tfrac{1}{2}\int_{x\in V}\{|\hat{f}_A(x)-f_A(x)|+|\hat{f}_B(x)-f_B(x)\}|dx \qquad (22)$$

If V is replaced by the whole space this becomes

$$\varepsilon\leq\varepsilon^{\star}+\tfrac{1}{2}\int_{\forall x}|\hat{f}_A(x)-f_A(x)|dx+\tfrac{1}{2}\int_{\forall x}|\hat{f}_B(x)-f_B(x)-f_B(x)|dx \qquad (23)$$

From (3) and (5) it follows that (23) can be written as

$$\varepsilon \leq \varepsilon^{\star} + e_A + e_B \qquad (24)$$

This is the basic formula of this paper.

From (14) another upperbound can be found by immediately replacing V by the whole space,

$$\varepsilon\leq\varepsilon^{\star}+\tfrac{1}{2}\int_{\forall x}|f_A(x)-f_B(x)|dx$$
$$\varepsilon\leq\tfrac{1}{2}\int_{\forall x}\{|f_A(x)-f_B(x)|+\min\{f_A(x),f_B(x)\}\}dx \qquad (25)$$

in which use has been made of (2). (25) is equivalent with

$$\varepsilon<\tfrac{1}{2}\int_{\forall x}\max\{f_A(x),\ f_B(x)\}dx$$
$$\varepsilon\leq\tfrac{1}{2}\int_{\forall x}\{f_A(x)+f_B(x)-\min\{f_A(x),f_B(x)\}\}dx$$
$$\varepsilon\leq 1-\varepsilon^{\star} \qquad (26)$$

Together with the obvious fact that $\varepsilon^{\star}\leq\varepsilon$ we yield

$$\varepsilon^{\star}\leq\varepsilon\leq\min\{(1-\varepsilon^{\star}),(e_A+e_B)\} \qquad (27)$$

For most practical problems (24) is a more stringent bound then (26).

We will consider two cases for the computation of e_A and e_B. First we adopt Hughes' model[4] of a general measurement space consisting of m cells, each with its own probability of occurence. This is a very general approach that allows all kinds of distributions. It appears therefore that e_A and e_B grow fast with m. A second case we investigate is that of normal distributions for $f_A(x)$ and $f_B(x)$. This leads, using a Monte Carlo procedure to more realistic figures.

Inequality (24) holds as it is given for a particular learning set. The error made by the discriminant function based on that learning set is expressed in the Bayes error and the estimation errors for the distributions. For an unknown case these errors are unknown, but the expectation of these errors can be computed, by a given class of distributions for the learning set.

$$E(\varepsilon)\leq\varepsilon^{\star}+E(e_A+e_B) \qquad (28)$$

The expectation of e_A+e_B can, as we will see, be expressed in the number of learning samples. This enables us to compute the sample size necessary for a certain accuracy.

General measurement space

We assume that x is an outcome in one of m cells with probabilities p_A^i and p_B^i (i=1,m) for class A and class B. For the estimation of both, p_A^i and p_B^i we assume to have n learning samples. The maximum likelihood estimates, indicated by $\hat{p}_A^i$ and $\hat{p}_B^i$ will be used. For (4) can be written:

$$e_A = 1 - \sum_{i=1}^{m} \min\{\hat{p}_A^i,\ p_A^i)\} \qquad (29)$$

Taking the expectation over all learning sets we yield:

$$E(e_A) = 1 - \sum_{i=1}^{m} E[\min\{\hat{p}_A^i,\ p_A^i\}] \qquad (30)$$

Define $y = (\hat{p}_A^i-p_A^i)\{p_A^i(1-p_A^i)/n\}^{\frac{1}{2}}$

so $E[\min\{\hat{p}_A^i,\ p_A^i\}] = E[\min\{y,0\}]\{p_A^i(1-p_A^i)/n\}^{\frac{1}{2}}+p_A^i$ (31)

If n is large enough $\hat{p}_A^i$ is normally distributed with expectation p_A^i and variance $p_A^i(1-p_A^i)/n$. y is then N(0,1) distributed. The first term in (31) can therefore be written as

$$E[\min\{y,0\}] = \int_{-\infty}^{0} y(2\pi)^{-\frac{1}{2}}\exp(-y^2/2)dy = -(2\pi)^{-\frac{1}{2}} \qquad (32)$$

because the integral is zero in the interval $0<y<\infty$. (30) yields with (31) and (32):

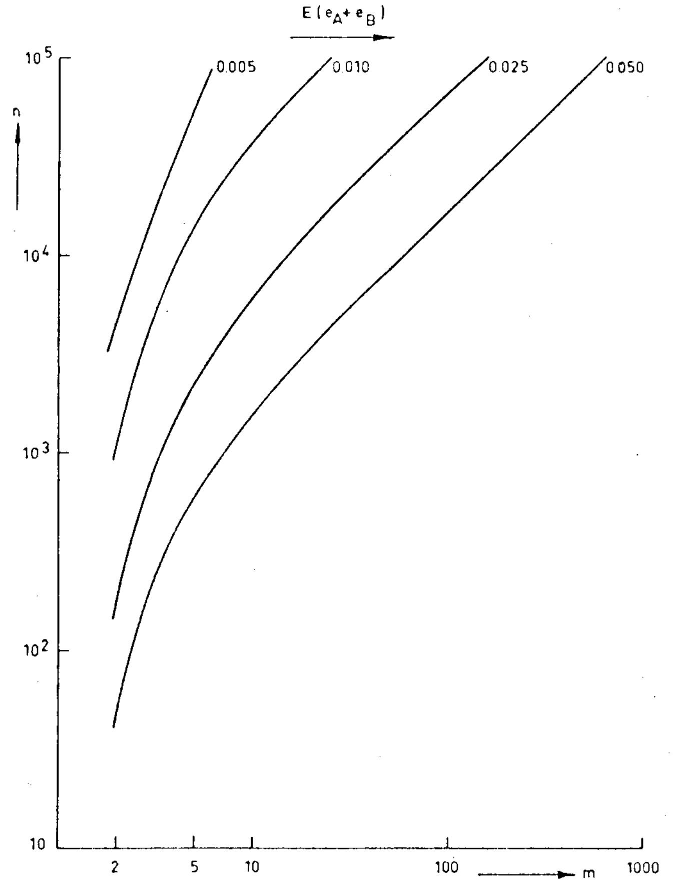

Fig. 1. The number of learning samples n as a function of the measurement complexity m for various values of $E(e_A + e_B)$, $\varepsilon^{\star} = 0.05$.

$$E(e_A) = \sum_{i=1}^{m} [(2\pi)^{-\frac{1}{2}}\{p_A^i(1-p_A^i)/n\}^{\frac{1}{2}}] \qquad (33)$$

because $\sum_{i=1}^{m} p_A^i=1$. In the same way can be derived

$$E(e_B) = \sum_{i=1}^{m} [(2\pi)^{-\frac{1}{2}}\{p_B^i(1-p_B^i)/n\}^{\frac{1}{2}}] \qquad (34)$$

In the appendix it is shown that $E(e_A+e_B)$ is maximum if for m/2 values of i

$$p_A^i = 2\varepsilon^\star/m,\ p_B^i = 2(1-\varepsilon^\star)/m \qquad (35)$$

and $p_A^i = 2(1-\varepsilon^\star)/m,\ p_B^i = 2\varepsilon^\star/m$ (36)

for the other m/2 values of i, (m even). We can write therefore for (28), using (33)-(36).

$$E(\varepsilon) \leq \varepsilon^\star + (2\pi n)^{-\frac{1}{2}} [\{2\varepsilon^\star(1-2\varepsilon^\star/m)/m\}^{\frac{1}{2}} + \\ + 2(1-\varepsilon^\star)(1-2(1-\varepsilon^\star)/m)/m\}^{\frac{1}{2}}] \qquad (37)$$

In the figures 1, 2 and 3 n is given as a function of m for constant values of $E(e_A+e_B)$ by $\varepsilon^\star$=0.2, 0.1 and 0.5 respectively. These curves should be interpreted as follows. If the Bayes error $\varepsilon^\star$ (infinite sample case) equals 0.1 and an additional error of 0.1 is allowed, due to the use of estimated distributions, then fig. 2 gives the number of learning samples as a function of the measurement complexity. The indicated sample size guarantees that the demanded accuracy is reached in expectation. Because the Bayes error $\varepsilon^\star$ is only rarely known in practice, the given curves serve mainly as an impression of the accuracy as a function of measurement complexity and sample size.

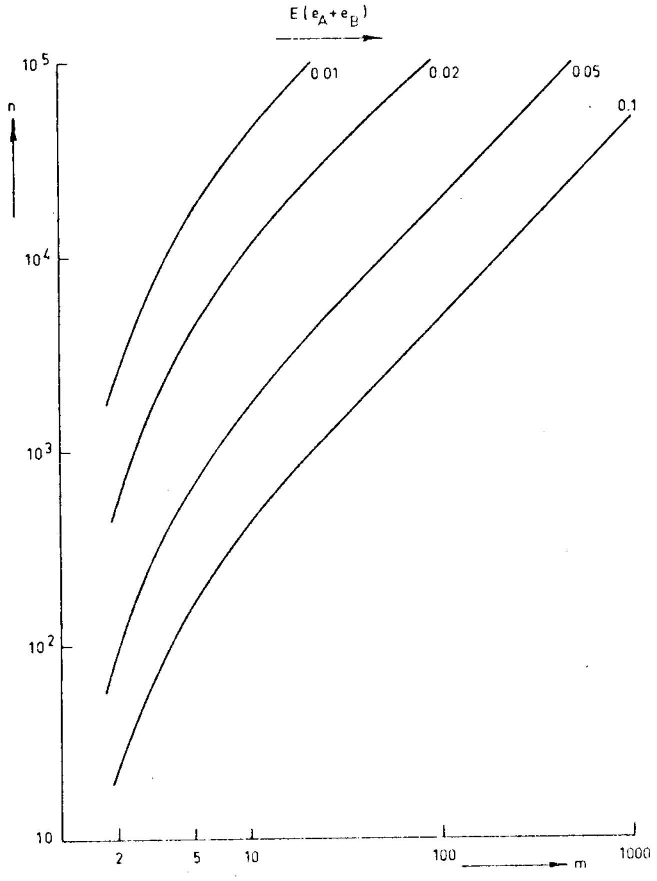

Fig. 2. The number of learning samples n as a function of the measurement complexity m for various values $E(e_A + e_B)$, $\varepsilon^\star$ = 0.10

The resulting numbers of learning samples are very large, and in many practical problems not available. Besides they are in respect to many practical results extremely pessimistic. This is caused by the very general approach in which nearly every kind of distribution is allowed. In the next section we will therefore restrict ourselves to normal distributions.

Normal distributions

In order to find $E(e_A+e_B)$ for k-dimensional normal distributions we consider the expectation of an estimation error e for that case. Using (4) we yield

$$E(e) = 1-E \int_{\forall x} \min\{f(x,\mu,\Sigma),f(x,\hat{\mu},\hat{\Sigma})\}dx \qquad (38)$$

in which $f(x,\mu,\Sigma)$ is the normal density function with expectation μ and covariance matrix Σ. The estimate of f is found using the maximum likelihood estimates $\hat{\mu}$ and $\hat{\Sigma}$ for μ and Σ. Any linear transformation of the x-space does not change the integral in (38) if the parameters of the distribution are transformed adequately. We may write therefore:

$$E(e) = 1-E[\int_{\forall x} \min\{f(x,0,I);f(x,\hat{\mu}',\hat{\Sigma}')\}dx] \qquad (39)$$

in which $\hat{\mu}'$ and $\hat{\Sigma}'$ are the transformations of $\hat{\mu}$ and $\hat{\Sigma}$.

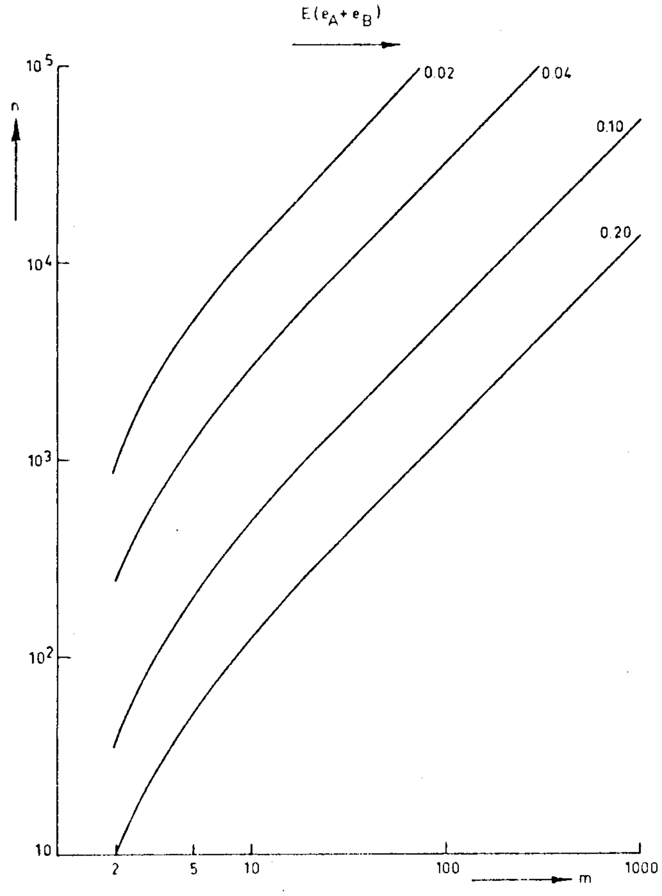

Fig. 3. The number of learning samples n as a function of the measurement complexity m for various values of $E(e_A + e_B)$, $\varepsilon^\star$ = 0.20

They can, however, also be interpreted as the maximum likelihood estimates of the parameters of the distribution f(x,0,I). (I is the unity matrix). We will not give the proof here, which is straight forward. The result of this reasoning is that E(e) is independent of μ and Σ, it only depends upon the dimensionality k and the number of learning samples n.

We computed E(e) as a function of n and k using Monte Carlo procedures. The integral of the minimum in (39) was found using a method described in [6] in which 50 points for each distribution were used. The expectation was obtained by averaging the results of 200 randomly chosen learning sets of size n. The accuracy of this method can be found by computing the standard deviation of those 200 results. In fig. 4 E(e), estimated in this way is shown as a function of n and k. The standard deviation in the averaged value is for the worst case about 0.007.

From fig. 4 n can be computed as a function of k for constant values of E(e). The result is shown in fig. 5. The values of n can be interpreted as that number of learning samples that guarantees that, in expectation, the contribution of the estimation error of some density function to the discriminant error is less than E(e). For a two class problem the values of E(e) should be multiplied by two in order to find $E(e_A+e_B)$.

One may wonder how valuable the numbers given in fig. 5 are in practice. In order to get an impression of that we did a number of experiments. If γ is defined by:

$$\varepsilon = \varepsilon^{\star} + \gamma(e_A+e_B) \qquad (40)$$

we computed for a number of classification problems with randomly chosen learning sets, ε, $\varepsilon^{\star}$, e_A and e_B. The resulting values of γ appeared very often to be less than 0.2. An example is given in table I where the results are presented of a two dimensional example. The distributions of the classes A and B were both independent. A with mean (0,0) and variances 1 and 1, B with mean (μ,0) and, variances v and 1. For each value of μ and v ten learning sets were chosen at random resulting in ten values of γ. In the table the mean value of γ and the number of times γ was larger than 0.15 or 0.20 is given. The results are given for n=20, and n=50.

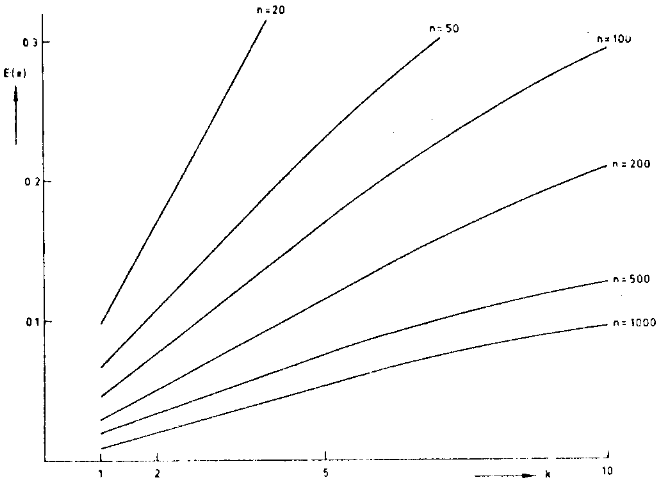

Fig. 4. The expected estimation error of a normal distribution as a function of dimensionality k and sample size n.

In order to be sure that there exists a bound such as γ=0.2 or some other value, much more experiments have to be done. Our experiments just showed that in a number of cases the accuracy is a factor five or more better than as determined from fig. 5.

μ	v	$\varepsilon^{\star}$	n = 20			n = 50		
			1)	2)	3)	1)	2)	3)
0	2	0.42	0.09	2	2	0.02	0	0
0.5	2	0.36	0.09	3	1	0.04	0	0
1.0	2	0.26	0.10	2	0	0.03	0	0
2.0	2	0.11	0.15	0	0	0.01	0	0
0	6	0.30	0.16	5	5	0.04	1	0
0.5	6	0.29	0.08	2	0	0.00	0	0
1.0	6	0.26	0.07	0	0	0.01	0	0
2.0	6	0.17	0.05	1	0	0.02	0	0
0	20	0.19	0.03	0	0	0.00	0	0
0.5	20	0.19	0.10	0	0	0.01	0	0
1.0	20	0.18	0.09	2	0	0.03	0	0
2.0	20	0.16	0.05	0	0	0.03	0	0

Table I

Results of a number of two dimensional experiments, each repeated for ten different learning sets and for sample sizes of 20 and 50 samples for each of the two classes. One distribution had a mean (0,0) and variances of 1 and 1, the other had a mean of (μ,0) and variances of v and 1. Both distributions were assumed to be independent.

1) mean value of γ in ten experiments
2) number of times γ > 0.15
3) number of times γ > 0.20

Multi-class case

The essence of the proof given for (24) is that each error made in the estimation of the density functions can result in an equal increase of the discriminant error. This causes for the multi-class case the following upperbound:

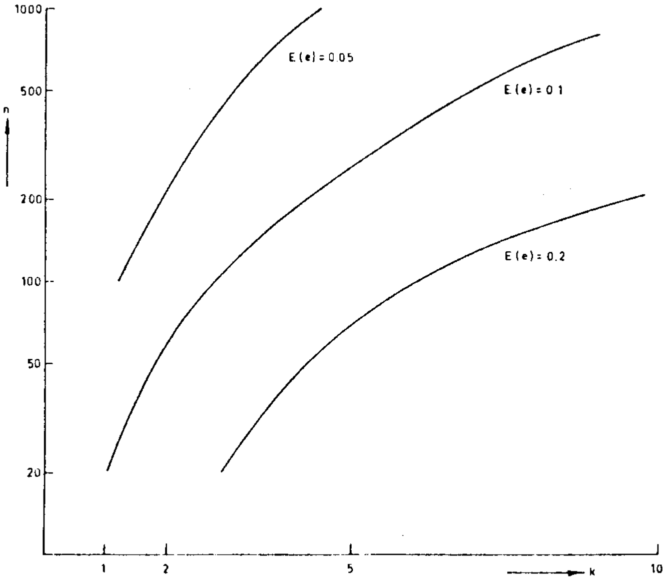

Fig. 5. The number of learning samples n as a function of the dimensionality k and the expected estimation error E(e) for normal distributions.

$$\varepsilon \leq \varepsilon^{\star} + \sum_{i=1}^{\ell} e_i \qquad (40)$$

ℓ is the number of classes, e_i is defined by (3). The general proof of (40) will not be given here.

An estimation error of a density function will only result in an increase of the discriminant error if that estimation error results in a different classification in a certain area. In a multi-class problem, however, several densities can be estimated erronously, but in each point only two of those errors can result in a different classification, the correct one and the one that takes over. The upperbound given in (40) is because of this reason in a multiclass problem weaker than in a two class problem. The results given in the previous sections can, nevertheless, be used in any ℓ-class problem. The values for n for the various cases can easily be derived from the given figures.

Summary

In this paper an upperbound is presented for the discriminant error. Because this bound is sample size dependent, it is possible to compute the guaranteed accuracy, in terms of expected errors, as a function of the number of learning samples. This has been worked out for two examples.

If we adopt Hughes'[5] model of a general measurement space, a rather conservative estimate is obtained, due to the very general approach. In case of normal distributions more realistic figures were obtained. These can even be improved by estimating to what extend the upperbound is approached in practice.

Acknowledgement

The author wishes to express his gratitude to prof. I.T. Young of M.I.T. for stimulating discussions and useful suggestions during his stay at Delft University of Technology.

Appendix

We will prove here that

$$\sum_{i=1}^{m} \{p_A^i(1-p_A^i)\}^{\frac{1}{2}} + \sum_{i=1}^{m} \{p_B^i(1-p_B^i)\}^{\frac{1}{2}} \qquad (41)$$

is maximum if

$$p_A^i = 2\varepsilon^{\star}/m, \; p_B^i = 2(1-\varepsilon^{\star})/m$$

in m/2 points, and

$$p_A^i = 2(1-\varepsilon^{\star})/m, \; p_B^i = 2\varepsilon^{\star}/m$$

in the other m/2 points.
As constraints for the maximization of (41) we have

$$\sum_{i=1}^{m} p_A^i = 1 \qquad (42)$$

$$\sum_{i=1}^{m} p_B^i = 1 \qquad (43)$$

$$\tfrac{1}{2} \sum_{i=1}^{m} \min\{p_A^i, p_B^i\} = \varepsilon^{\star} \qquad (44)$$

Using the Lagrange multiplier method we yield

$$\text{max:} \sum_{i=1}^{m} \{p_A^i(1-p_A^i)\}^{\frac{1}{2}} + \sum_{i=1}^{m} \{p_B^i(1-p_B^i)\}^{\frac{1}{2}} + \lambda(\sum_{i=1}^{m} p_A^i - 1) + $$
$$+ \mu(\sum_{i=1}^{m} p_B^i - 1) + \nu(\tfrac{1}{2} \sum_{i=1}^{m} \min\{p_A^i, p_B^i - \varepsilon^{\star}) \qquad (45)$$

Suppose $p_A^i < p_B^i$. The derivative to p_A^i is equal to zero if

$$\tfrac{1}{2}\{p_A^i(1-p_A^i)\}^{-\frac{1}{2}}(1-2p_A^i)+\lambda+\tfrac{1}{2}) = 0 \qquad (46)$$

The derivative to p_B^i is equal to zero if

$$\tfrac{1}{2}\{p_B^i(1-p_B^i)\}^{-\frac{1}{2}}(1-2p_B^2)+\mu = 0 \qquad (47)$$

(46) and (47) give for all i, with $p_A^i < p_B^i$ the same solution. In the same way can be proved that all i with $p_B^i \leq p_A^i$ yield the same solution. Suppose that for m_1 values of i, $p_A^i < p_B^i$ with solution $p_A^i = p_A^1$, $p_B^i = p_B^1$ and that for $m_2 (m_2 = m - m_1)$ values of i, $p_B^i \leq p_A^i$ with solution $p_A^i = p_A^2$, $p_B^i = p_B^2$. By using (41)-(44) we now get
max:

$$m_1\{p_A^1(1-p_A^1)\}^{\frac{1}{2}} + m_2\{p_A^2(1-p_A^2)\}^{\frac{1}{2}} + m_1\{p_B^1(1-p_B^1)\}^{\frac{1}{2}} + m_2\{p_B^2(1-p_B^2)\}^{\frac{1}{2}} \qquad (48)$$

with constraints

$$m_1 p_A^1 + m_2 p_A^2 = 1 \qquad (49)$$

$$m_1 p_B^1 + m_2 p_B^2 = 1 \qquad (50)$$

$$\tfrac{1}{2} m_1 p_A^1 + \tfrac{1}{2} m_2 p_B^2 = \varepsilon^{\star} \qquad (51)$$

With (49) p_A^2 can be expressed in p_A^1, with (51) p_B^2 can be expressed in p_{A_1} and with this result and (50) p_B^1 can be expressed in p_A^1. If we write for m_2 $m-m_1$ we now yield for (48):

$$\text{max: } m_1\{p_A^1(1-p_A^1)\}^{\frac{1}{2}} + \{(1-m_1 p_A^1)(m-m_1-1+m_1 p_A^1)\}^{\frac{1}{2}} +$$
$$+ \{(1-2\varepsilon^{\star}+m_1 p_A^1)(m_1-1+2\varepsilon^{\star}-m_1 p_A^1)\}^{\frac{1}{2}} +$$
$$+ \{(2\varepsilon^{\star}-m_1 p_A^1)(m-m_1-2\varepsilon^{\star}+m_1 p_A^1)\}^{\frac{1}{2}} \qquad (52)$$

This is a sum of roots of quadratic functions in p_A^1. The derivative is a sum of monotonous decreasing functions and has therefore only one zerocrossing. According to the same reasoning there exists only one solution for m_1. The proof is completed by the computation of the two derivatives and the substitution of the solution given in the start of this appendix, which implies $m_1 = m/2$. Using (49)-(51) p_A^2, p_B^1 and p_B^2 can be computed. We will omit the calculations of this final part of the proof because of its spacious character.

References

1 K. Fukanaga, "Introduction to Statistical Pattern Recognition", Academic Press, 1972.

2 T.M. Cover, "Geometrical and Statistical Properties of systems of linear inequalities with applications in pattern recognition", IEEE Trans. Electronic. Comp., vol. EC-14, pp. 326-334, June 1965.

3 D.H. Foley, "Considerations of Sample and Feature Size", IEEE Trans. Inform. Theory, vol. IT-18, pp. 618-626, Sept. 1972.

4 G.F. Hughes, "On the Mean Accuracy of Statistical Pattern Recognizers", IEEE Trans. Inform. Theory, vol. IT-14, no. 1, pp. 55-63, January 1968.

5 K. Abend, T.J. Harley, B. Chandrasekaran, G.F. Hughes, "Comments "On the mean accuracy of statistical Pattern Recognizers"", IEEE Trans. Inform. Theory, vol. IT-15, pp. 420-423, May 1969.

6 R.P.W. Duin, "A criterion for the smoothing parameter for Parzen estimators of probability density functions", Internal report Pattern Recognition Group, Department of Applied Physics, Delft University of Technology, Sept. 1975.