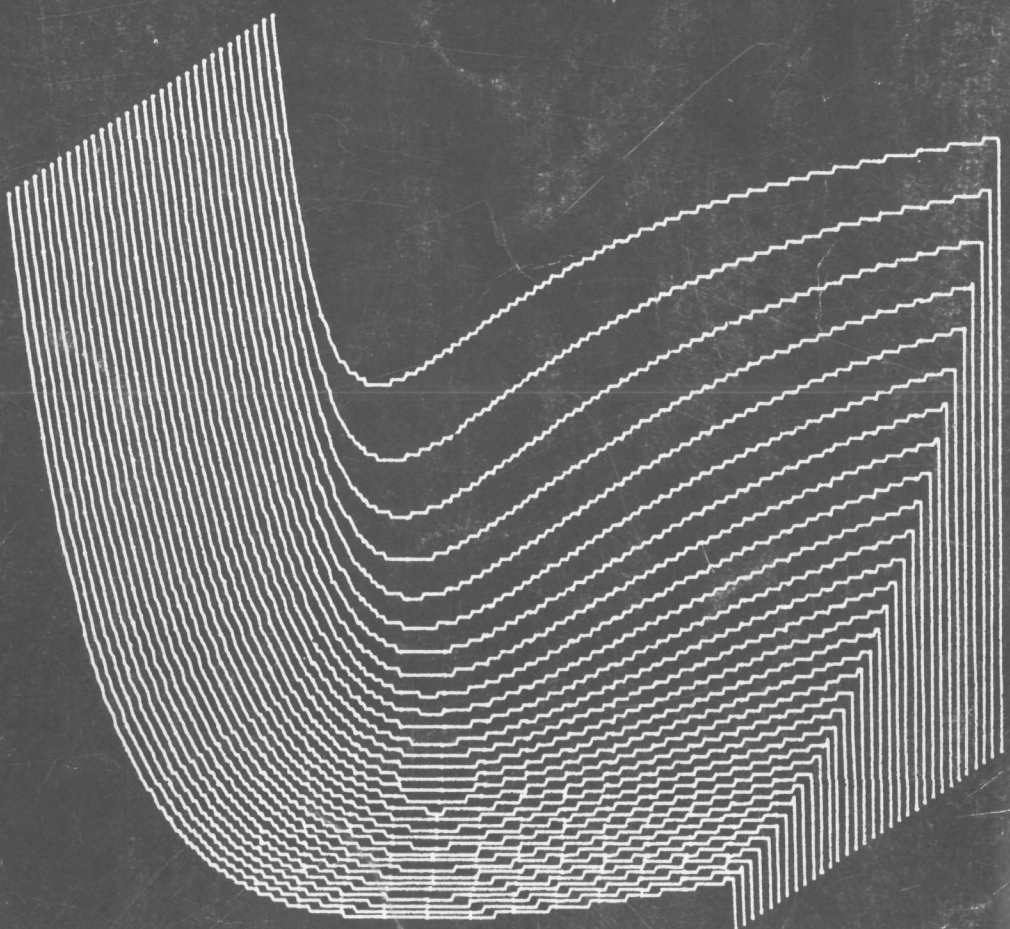
ON THE ACCURACY OF STATISTICAL PATTERN RECOGNIZERS

R. P. W. DUIN



ON THE ACCURACY OF STATISTICAL PATTERN RECOGNIZERS

ISBN 90 6231 052 4 soft-bound edition
ISBN 90 6231 053 2 hard-bound edition

ON THE ACCURACY OF STATISTICAL PATTERN RECOGNIZERS

PROEFSCHRIFT

ter verkrijging van de graad van doctor in de technische wetenschappen aan de Technische Hogeschool Delft, op gezag van de rector magnificus Prof. Ir. L. Huisman, voor een commissie aangewezen door het college van dekanen te verdedigen op woensdag 14 juni 1978 te 14.00 uur

door

ROBERT PIETER WILHELM DUIN
natuurkundig ingenieur,
geboren te Maasniel

1978
Dutch Efficiency Bureau – Pijnacker

Dit proefschrift is goedgekeurd door de promotoren

PROF.IR. J.W. SIEBEN

PROF.DR.IR. C.J.D.M. VERHAGEN

aan mijn ouders
aan inge

CONTENTS

SUMMARY

The accuracy of a statistical pattern recognizer depends upon the intrinsic overlap of the class distributions and the estimation error due to the finite size of the set of learning objects. The classification error, defined as the probability of error in classifying new objects, is used as a measure for the accuracy. The classification error depends on the characteristics of the features chosen, the number of features, the size of the learning set, and on the procedure and the statistical model used for the estimation of the discriminant function. A number of these items are influenced or determined by a priori knowledge.

In order to be able to study the effects of feature size and sample size for given class distributions the expected classification error is investigated, which is the expectation of the classification error over the population of learning sets. More general aspects are studied by using the mean classification error, which is defined as the expectation of the expected classification error over a family of class ditributions.

For the expected classification error an upper bound is found expressed in the intrinsic overlap of the class distributions and the expectations of the estimation errors due to using a finite sample size for the estimation of the class distributions. The upper bound is applied to the case of a discrete measurement space and to the case of normally distributed classes. For these cases sample sizes are computed which guarantee, in expectation, a specified classification accuracy.

The expected classification errors for a histogram and a Parzen estimate are compared for the case of a one dimensional nonparametric density estimation. The errors in the density estimates and the classification errors of Parzen estimates using normal and uniform kernels are compared in a multi dimensional example.

If the classification error is studied for a finite sample size as function

of the feature size a peaking phenomenon may be observed: after a certain feature size the classification error starts to increase and approaches the a priori error if the feature size grows to infinity. It is shown that the mean classification error does not peak if the right estimators are used. This leads to a better understanding of the causes of peaking: bad features and bad estimators. Sufficient conditions could be formulated for the feature distributions and for the estimators under which peaking of the mean classification error is avoided. Also a better understanding of the discrete measurement space introduced by Hughes [24] is reached. It appeares that peaking in that space is caused by the specific properties of that space.

The influence of the complexity of the statistical model (i.e. the number of parameters in that model) on the classification error is illustrated by a two dimensional example of two normally distributed classes. It appears that the expected classification error as a function of the model complexity may cause peaking too.

The choice of good features, the ranking of the features, the determination of their optimal number, the choice of good estimators and of a statistical model all make it necessary to have some a priori knowledge of the distributions of the possible features. Without such knowledge it is hardly possible, because of the peaking phenomenon, to obtain good classification results using a finite learning set.

SAMENVATTING

De nauwkeurigheid van een statistische procedure voor patroonherkenning hangt af van de intrinsieke overlap van de verdelingen van de klassen en van de schattingsfout als gevolg van de eindige grootte van de verzameling leerobjecten. De klassificatiefout, gedefinieerd als de foutkans bij het klassificeren van nieuwe objecten wordt gebruikt als een maat voor de nauwkeurigheid. De klassificatiefout is afhankelijk van de eigenschappen van de gekozen kenmerken, van het aantal kenmerken, van de grootte van de verzameling leerobjecten en van de procedure en het statistisch model die gebruikt zijn bij de schatting van de scheidingsfunctie. Een aantal van deze grootheden wordt beïnvloed of bepaald door a priori kennis.

Teneinde in staat te zijn bij gegeven verdelingen van de klassen de invloed van het aantal kenmerken en het aantal leerobjecten na te gaan is de verwachte klassificatiefout berekend. Dit is de verwachting van de klassificatiefout over de populatie van verzamelingen leerobjecten. Meer algemene aspecten zijn bestudeerd door gebruik te maken van de gemiddelde klassificatiefout, gedefinieerd als de verwachting van de verwachte klassificatiefout over een familie van klasseverdelingen.

Voor de verwachte klassificatiefout is een bovengrens afgeleid die is uitgedrukt in de intrinsieke overlap van de verdelingen van de klassen en de verwachtingen van de schattingsfouten als gevolg van het gebruik van een eindige verzameling leerobjecten voor het schatten van de klasseverdelingen. De bovengrens is toegepast voor het geval van een discrete meetruimte en voor het geval van normaal verdeelde klassen. Voor deze gevallen zijn de aantallen leerobjecten berekend die een gegeven nauwkeurigheid garanderen voor het verwachte klassificatieresultaat.

De verwachte schattingsfouten voor een histogram en voor een parzenschatting zijn vergeleken voor het geval van een ééndimensionale niet-parametrische dichtheidschatting. De fouten in de dichtheidschattingen en de klassificatiefouten zijn voor een meerdimensionaal voorbeeld vergeleken bij

gebruik van parzenschatters met uniforme en normale kernels.

Als in geval van een eindige verzameling leerobjecten de klassificatiefout wordt bestudeerd als functie van het aantal kenmerken, dan kan een 'piekeffect' worden waargenomen: na een zeker aantal kenmerken begint de klassificatiefout toe te nemen en nadert naar de a priori fout als het aantal kenmerken naar oneindig gaat. Er is aangetoond dat bij gebruik van de juiste schatters de gemiddelde klassificatiefout dit piekeffect niet vertoont. Dit leidt tot een beter begrip voor de oorzaken van dit piekeffect: slechte kenmerken en slechte schatters. Voldoende voorwaarden zijn afgeleid m.b.t. de verdeling van de kenmerken en de gebruikte schatters waaronder voor de gemiddelde klassificatiefout het optreden van dit piekeffect wordt voorkomen. Tevens wordt hierdoor een beter begrip bereikt voor de discrete meetruimte zoals die is gebruikt door Hughes [24]. Het blijkt dat in die ruimte het piekeffect wordt veroorzaakt door de speciale eigenschappen van die ruimte.

In een tweedimensionaal voorbeeld van twee normaal verdeelde klassen wordt de invloed geïllustreerd van de complexiteit van het statistisch model (dat is het aantal parameters van het model) dat wordt gebruikt bij de schatting van de dichtheden van de klassen. Het blijkt dat het piekeffect eveneens kan optreden voor de verwachte klassificatiefout als functie van de modelcomplexiteit.

De keuze van goede kenmerken, hun rangschikking, de bepaling van het optimale aantal kenmerken, de keuze van goede schatters en van het statistisch model maken het te zamen noodzakelijk enige a priori kennis te hebben over de verdeling van de kenmerken. Zonder dergelijke kennis is het nauwelijks mogelijk, met het oog op het piekeffect, tot goede klassificatieresultaten te komen bij gebruik van een eindige verzameling leerobjecten.

Chapter 1
INTRODUCTION

The goal of statistical pattern recognition is to analyse the patterns that may be present in a set of objects in terms of measurements on these objects using statistics and a priori knowledge. This is often done in order to be able to classify the objects. Objects in this context are broadly interpreted. All that can be described by a set of measurements, such as a movement made by an arm, a sunny day, a human chromosome, a polluted river or a line on a picture is included.

Measurements or transformations thereof, that may be useful for the description of a pattern in relation to other patterns are called features. We will assume that a set of those features as well as useful statistical models for the description of patterns are available as a priori knowledge. In this thesis we will restrict ourselves to the so called case of supervised learning. In this case sets of objects are given, in which previously certain patterns have been identified by man or by other means. A decision rule has to be constructed to decide between patterns in those sets. A set of objects with a common pattern is called a class. A learning set, which is a set of objects with known classification, is assumed to be available for the construction of the decision rule, which is a discriminant function on the features between the classes.

For the accuracy of a statistical pattern recognizer several measures will be introduced later on. They are all related to the probability of error in classifying new objects. This probability of error will further be called the *classification error*, or just error. It depends on the characteristics of the features, the number of features, the size of the learning set, and the statistical model and the estimation procedure used for the determination of the discriminant function. A number of these quantities are influenced or determined by a priori knowledge.

In the next five chapters we will present some discussions on each of these

subjects. Because they are highly interrelated, the order presented is sometimes arbitrary. Theoretical as well as experimental results using generated data will be given. Some are not yet mentioned in literature, some are already published by the author, others are well known and only given for illustration and comparison. Similar problems are studied by Raudys [36] but in a less general way and applied to more special types of discriminant functions. A general discussion of some aspects of the problem is given by Kanal and Chandrasekaran [25].

Because our goal is merely to study a number of general effects and relations and not to give a complete guide for practical purposes we will make some additional assumptions that simplify the notation and avoid some additional problems not essential to this research. We will restrict ourselves almost entirely to the case of two classes A and B, given by learning sets of the same size m. The a priori probabilities of the classes will be c for class A and 1-c for class B and are assumed to be known.

Before presenting a short introduction to the other chapters we will define the main points of the notation and terminology. An arbitrary object, characterized by k features, will be denoted by the k-dimensional vector $\underline{x} = (x_1, x_2, x_3, \ldots x_k)$, in which $x_j (j=1,k)$ is a feature value. The probability density function for class $\ell (\ell=A,B)$ is written as $f_\ell(\underline{x})$ or as $f_\ell(\underline{x}|\underline{\theta}_\ell)$. The parameter vector $\underline{\theta}_\ell$ contains all parameters of the function $f_\ell(\cdot)$ introduced by the choice of the features. A different feature set or a different feature ranking will cause a different functional form of $f_\ell(\cdot)$ or just a different value of $\underline{\theta}_\ell$. Whenever we write $f_\ell(\underline{x}|\theta_\ell)$ we assume that a particular feature choice is made and that the parameter vector has the value $\underline{\theta}_\ell$. For simplicity the vector $\underline{\theta}$, defined as $\underline{\theta} = (\underline{\theta}_A : \underline{\theta}_B)$ will be used sometimes.

In this thesis the so called Bayes strategy will be used for finding a discriminant function. This implies the minimization of the expected costs. Everywhere will be assumed that the costs of a correct classification are zero and the costs involved with an erroneous classification are equal for the two classes. Under these restrictions the Bayes strategy is equivalent to minimizing the probability of misclassification (see Fukunaga [23]). The discriminant function becomes in that case

$$S(\underline{x}) = c\, f_A(\underline{x}) - (1-c)\, f_B(\underline{x}) \qquad (1.1)$$

which classifies as

if $S(\underline{x}) > 0$ then $\underline{x} \in$ class A

if $S(\underline{x}) = 0$ then $\underline{x} \in$ class A or $\underline{x} \in$ class B (1.2)

if $S(\underline{x}) < 0$ then $\underline{x} \in$ class B.

The boundary case $S(\underline{x}) = 0$ will be assigned to class A arbitrarily.

Instead of $S(\underline{x})$ the discriminant function $R(\underline{x})$ will be used sometimes

$$R(\underline{x}) = \log\{c\ f_A(\underline{x})\} - \log\{(1-c)\ f_B(\underline{x})\} \tag{1.3}$$

This function classifies in the same way as (1.2). When $S(\underline{x})$ or $R(\underline{x})$ is studied as a function of $\underline{\theta}$ they will be written as $S(\underline{x},\underline{\theta})$ or $R(\underline{x},\underline{\theta})$. In this context $f_A(\underline{x}|\underline{\theta}_A)$ and $f_B(\underline{x}|\underline{\theta}_B)$ are written for $f_A(\underline{x})$ and $f_B(\underline{x})$.

The classification error ε^* made by classifying with $S(\underline{x})$ given by (1.1) is

$$\varepsilon^* = c\ \text{Prob}\ (S(\underline{x}) < 0 \mid \underline{x} \in A) + (1-c)\ \text{Prob}\ (S(\underline{x}) \geqslant 0 \mid \underline{x} \in B) \tag{1.4}$$

which is equivalent to

$$\varepsilon^* = c \int_{S(\underline{x})<0} f_A(\underline{x})\ d\underline{x} + (1-c) \int_{S(\underline{x})\geq 0} f_B(\underline{x})\ d\underline{x} \tag{1.5}$$

The same is true when $R(\underline{x})$, given by (1.3), is used for $S(\underline{x})$.

Whenever possible, the compact notation of (1.5) will be used for the multi-dimensional integration in which $d\underline{x}$ stands for dx_1, dx_2, dx_3, ... dx_k. In the general discussions we will assume that $\underline{x}$ is a multidimensional continuous variable. By using summations instead of integrals the results apply also to the discrete case. Because the discriminant function ((1.1) or (1.3)) is the optimal one, the error ε^* is minimum. It is called the *Bayes error*.

From (1.1) and (1.2) it can be understood that ε^* can also be written as

$$\varepsilon^* = \int_{\underline{x}} \min\{c\ f_A(\underline{x}),\ (1-c)\ f_B(\underline{x})\}d\underline{x} \tag{1.6}$$

Note that $\varepsilon^* \leqslant \min\{c,1-c\}$, in which the equal sign applies to $f_A(\underline{x}) \equiv f_B(\underline{x})$. When $f_A(\underline{x})$ and $f_B(\underline{x})$ are unknown they have to be estimated using a learning set. Such a set will be denoted by $\chi = \chi_A \cup \chi_B$ in which $\chi_\ell = \{\underline{x}_\ell^1, \underline{x}_\ell^2, \underline{x}_\ell^3, \ldots \underline{x}_\ell^m\}$, ℓ = A,B. An object $\underline{x}_\ell^i$ is learning object number i of class ℓ. The learning objects are assumed to be selected independently according to the densities $f_A(\underline{x})$ and $f_B(\underline{x})$. The number of learning objects m is often called the sample

size. Suppose $f_A(\underline{x})$ is estimated by $\hat{f}_A(\underline{x})$ and $f_B(\underline{x})$ by $\hat{f}_B(\underline{x})$. An estimate of $S(\underline{x})$ is

$$\hat{S}(\underline{x}) = c\ \hat{f}_A(\underline{x}) - (1-c)\ \hat{f}_B(\underline{x}) \tag{1.7}$$

The classification error ε made by classifying with $\hat{S}(\underline{x})$ based on a given learning set χ is given by

$$\varepsilon = c\ \text{Prob}\ (\hat{S}(\underline{x}) < 0 \mid \underline{x} \in A,\chi) + (1-c)\ \text{Prob}\ (\hat{S}(\underline{x}) \geqslant 0 \mid \underline{x} \in B,\chi) \tag{1.8}$$

which is equivalent to

$$\varepsilon = c \int_{\hat{S}(\underline{x})<0} f_A(\underline{x})\ d\underline{x} + (1-c) \int_{\hat{S}(\underline{x})\geq 0} f_B(\underline{x})\ d\underline{x} \tag{1.9}$$

The classification error ε can only be computed by (1.8) and (1.9) if the density functions $f_A(\underline{x})$ and $f_B(\underline{x})$ are known. This is the case during simulations in which special choices are made for these densities. A value of ε is the result of a single experiment by which one learning set χ is generated, $\hat{S}(\underline{x})$ is estimated and (1.9) is computed. Such a value of ε can be considered as a random variable in respect to the choice of χ and is for that reason not very suitable as a measure for the accuracy of the discriminant procedure used. For that reason the *expected classification error*

$$\bar{\varepsilon} = E_\chi(e) \tag{1.10}$$

is a feasible quantity to study in relation to the expected accuracy of a statistical pattern recognizer. This error gives the expected performance in a single problem.

In order to investigate more general aspects of the classification error and to be able to make statements which are more problem independent $\bar{\varepsilon}$ will be averaged over a class of problems. For mathematical convenience we will restrict ourselves to those classes of problems which can be generated by a distribution over $\underline{\theta}_A$ and $\underline{\theta}_B$ for given functional forms of $f_A(\cdot)$ and $f_B(\cdot)$ and given feature size k. We therefore introduce the *mean classification error*

$$\tilde{\varepsilon} = E_{\underline{\theta}}\ E_\chi(\varepsilon) \tag{1.11}$$

in which E_θ is the expectation over the distribution of $\underline{\theta}_A$ and $\underline{\theta}_B$ that defines the class of problems of interest. The mean classification error $\tilde{\varepsilon}$ can be treated as a measure for the accuracy of a statistical pattern recognizer if it is studied in relation to a class of problems. A short introduction to the literature and the types of problems we will deal with is given by Duda and Hart [11,(sections 3.8 - 3.10)].

Most of the results depend upon the estimators used for $\hat{f}_A(\underline{x})$, $\hat{f}_B(\underline{x})$, $\hat{S}(\underline{x})$ and $\hat{R}(\underline{x})$. Most estimators that are needed here will be presented in chapter 2. Their effect upon ε, $\bar{\varepsilon}$ and $\tilde{\varepsilon}$ is shown by examples.

In chapter 3 especially the effect of the sample size upon the classification error is studied. For some special distributions curves are given for the expected error as a function of sample size and feature size. For other cases an upper bound is given for the expected error.

Feature size and sample size are closely related. If the number of features increases, the sample size necessary for a constant ε may increase, decrease or remain equal. This depends upon the characteristics of the new features. It can therefore occur that the expected error increases by increasing feature size and constant sample size. This effect is called peaking. It was first studied by Hughes [24] and later by Kanal, Chandrasekaran et al. [1], [5], [7], [25]. In chapter 4 feature size considerations are given with emphasis on the peaking effect.

In chapter 5 some examples are given of the effect of the choice of the statistical model used for the density function estimates $\hat{f}_A(\underline{x})$ and $\hat{f}_B(\underline{x})$ upon the classification error. It appears that a wrong model can result in a smaller error than the right model, especially in the case of small sample size.

In chapter 6 some remarks are made on the influence and use of a priori knowledge on the classification error. Especially the necessity of some knowledge about useful features is emphasized. If no such knowledge is available hardly any statistical pattern recognition is possible. Some results are discussed in the light of epistomology.

The main conclusions are summarized and discussed in chapter 7.

Chapter 2
ESTIMATION PROCEDURES

An unknown density function $f_\ell(\underline{x})$ can be estimated in various ways from a randomly chosen learning set. The most general techniques are the non-parametric ones such as the use of histograms or Parzen estimators. They demand very little knowledge of the functional form of $f_\ell(\underline{x})$ and are consistent under mild conditions (see for instance Patrick [32]). For our purposes they are not very well suited, because the computation of $\bar{\varepsilon}$ and $\tilde{\varepsilon}$ requires an integration over the learning set, which is only feasible in very simple situations and by using Monte Carlo procedures. In 3.4 some examples will be given.

If the functional form of $f_\ell(\underline{x})$ is known without the values of the parameters, a number of ways exist for the estimation of $f_\ell(\underline{x})$. Besides it is sometimes possible to estimate the discriminant function $S(\underline{x})$ directly, as will be shown below. In that case too, however, $\hat{S}(\underline{x})$ can be interpreted as being built up from estimates of the class density functions. The following three procedures for the estimation of the discriminant function

$$S(\underline{x}) = c\, f_A(\underline{x}|\underline{\theta}_A) - (1-c)\, f_B(\underline{x}|\underline{\theta}_B) \tag{2.1}$$

will be used in the next chapters.

1) First we will consider the *plug-in rule*. It is based on finding estimates $\hat{\underline{\theta}}_A$ for $\underline{\theta}_A$ and $\hat{\underline{\theta}}_B$ for $\underline{\theta}_B$ and simply 'plugging in' these estimates in (2.1).

$$\hat{S}^{(1)}(\underline{x}) = c\, f_A(\underline{x}|\hat{\underline{\theta}}_A) - (1-c)\, f_B(\underline{x}|\hat{\underline{\theta}}_B) \tag{2.2}$$

This implies that the density functions are estimated as

$$\hat{f}_\ell^{(1)}(\underline{x}) = f_\ell(\underline{x}|\hat{\underline{\theta}}_\ell) \quad (\ell = A,B) \tag{2.3}$$

The way the parameters are estimated is still open. We will often make use of *maximum likelihood estimates* for $\underline{\theta}_A$ and $\underline{\theta}_B$

$$\hat{\underline{\theta}}_\ell = \underset{\underline{\theta}_\ell}{\text{argmax}} \{ \prod_{i=1}^{m} f_\ell(\underline{x}_\ell^i | \underline{\theta}_\ell) \} \quad (\ell = A,B) \tag{2.4}$$

The argmax-function yields that value of $\underline{\theta}_\ell$ for which the argument is maximum.

2) The plug-in rule is very commonly used because it does not necessarily assume any knowledge on $\underline{\theta}$. This rule, however, is not optimal, as will be shown, if one deals with a class of problems with known distribution over $\underline{\theta}$. Using the Bayes rule it is possible to find the *a posteriori density* $g_\ell(\underline{\theta}_\ell | \chi_\ell)$ for the class parameters $\underline{\theta}$ using the *a priori density* $h_\ell(\underline{\theta}_\ell)$ (see Duda and Hart [11]).

$$g_{\underline{\theta}_\ell}(\underline{\theta}_\ell | \chi_\ell) = \frac{g_\ell(\chi_\ell | \underline{\theta}_\ell)\ h_\ell(\underline{\theta}_\ell)}{\int_{\underline{\theta}_\ell} g_\ell(\chi_\ell | \underline{\theta}_\ell)\ h_\ell(\underline{\theta}_\ell)\ d\underline{\theta}_\ell} \quad (\ell = A,B) \tag{2.5}$$

in which $h_\ell(\underline{\theta}_\ell)$ is the a priori density of $\underline{\theta}_\ell$ and

$$g_\ell(\chi_\ell | \underline{\theta}_\ell) = \prod_{i=1}^{m} f_\ell(\underline{x}_\ell^i | \underline{\theta}_\ell) \quad (\ell = A,B) \tag{2.6}$$

is the joint density of the learning objects of class ℓ. An estimate of $f_\ell(\underline{x} | \underline{\theta}_\ell)$ can now be found by estimating $\underline{\theta}_\ell$ from (2.5) by taking the expectation of $\underline{\theta}_\ell$ (called the *Bayes estimate* of $\underline{\theta}_\ell$) and using the plug-in rule. Another possibility is taking the expectation of $f_\ell(\underline{x} | \underline{\theta}_\ell)$ over $g_{\underline{\theta}_\ell}(\underline{\theta}_\ell | \chi_\ell)$ and obtaining the Bayes estimate of $f_\ell(\underline{x} | \underline{\theta}_\ell)$

$$\hat{f}_\ell^{(2)}(\underline{x}) = \int_{\underline{\theta}_\ell} f_\ell(\underline{x} | \underline{\theta}_\ell)\ g_{\underline{\theta}_\ell}(\underline{\theta}_\ell | \chi_\ell)\ d\underline{\theta}_\ell \quad (\ell = A,B) \tag{2.7}$$

In this way the following estimate of $S(\underline{x})$ is found

$$\hat{S}^{(2)}(\underline{x}) = c\ \hat{f}_A^{(2)}(\underline{x}) - (1-c)\ \hat{f}_B^{(2)}(\underline{x}) \tag{2.8}$$

We prefer the Bayes estimate of $f_\ell(\underline{x} | \underline{\theta}_\ell)$ to the Bayes estimate of $\underline{\theta}_\ell$ and using the plug-in rule because it is immediately related to the expectation of $S(\underline{x},\underline{\theta})$

over the a posteriori distributions for $\underline{\theta}_A$ and $\underline{\theta}_B$. In the case of binomially distributed and independent features these two estimates are identical. This follows straight forward from substitution of the densities in (2.6) and computing (2.5) and (2.7). It is caused by the fact that the parameter p of a binomial distribution is identical with the density for x=1 : p = f(1).

3) A third way of estimating the discriminant function is found by taking the expectation of $S(\underline{x},\underline{\theta})$ over the a posteriori distribution of $\underline{\theta}$.

$$\hat{S}^{(3)}(\underline{x}) = \int_{\underline{\theta}} S(\underline{x},\underline{\theta})\ g_{\underline{\theta}}(\underline{\theta}|\chi)\ d\underline{\theta} \tag{2.9}$$

where $g_{\underline{\theta}}(\underline{\theta}|\chi)$ is given by

$$g_{\underline{\theta}}(\underline{\theta}|\chi) = \frac{g(\chi|\underline{\theta})\ h(\underline{\theta})}{\int_{\underline{\theta}} g(\chi|\underline{\theta})\ h(\underline{\theta})\ d\underline{\theta}} \tag{2.10}$$

In (2.10) is

$$g(\chi|\underline{\theta}) = \prod_{i=1}^{m} \{f_A(\underline{x}_A^i|\underline{\theta}_A)\ f_B(\underline{x}_B^i|\underline{\theta}_B)\} \tag{2.11}$$

the joint density of all learning objects and $h(\underline{\theta})$ the density of the whole set of parameters $\underline{\theta}$. Substitution of (2.1) in (2.9) yields

$$\hat{S}^{(3)}(\underline{x}) = c \int_{\underline{\theta}} f_A(\underline{x}|\underline{\theta}_A)\ g_{\underline{\theta}}(\underline{\theta}|\chi)\ d\underline{\theta} - (1-c) \int_{\underline{\theta}} f_B(\underline{x}|\underline{\theta}_B)\ g_{\underline{\theta}}(\underline{\theta}|\chi)\ d\underline{\theta} \tag{2.12}$$

This can be written as

$$\hat{S}^{(3)}(\underline{x}) = c\ \hat{f}_A^{(3)}(\underline{x}) - (1-c)\ \hat{f}_B^{(3)}(\underline{x}) \tag{2.13}$$

with

$$\hat{f}_\ell^{(3)}(\underline{x}) = \int_{\underline{\theta}} f_\ell(\underline{x}|\underline{\theta}_\ell)\ g_{\underline{\theta}}(\underline{\theta}|\chi)\ d\underline{\theta} \qquad (\ell = A,B) \tag{2.14}$$

Note the difference between (2.7) where $\hat{f}^{(2)}(\underline{x})$ just depends upon χ_ℓ, the learning set of class ℓ, and (2.14) where $\hat{f}^{(3)}(\underline{x})$ depends upon the entire learning set χ. Note also that these two estimates become identical if

$$h(\underline{\theta}) = h_A(\underline{\theta}_A)\, h_B(\underline{\theta}_B) \tag{2.15}$$

which can easily be verified by substitution of (2.10) and (2.11) in (2.14) using (2.15). This method is equivalent with a method known as 'predictive diagnosis' in the medical statistical literature, e.g. see Aitchinson, Habbema and Kay [2].

We will give a simple example in order to illustrate the differences between the three types of estimates. The interesting point in this example is the difference in results for the three kinds of estimators in the multivariate case. For simplicity, however, the estimates will be given for the one dimensional case only.

Let a feature be binomially distributed for the classes A and B.

$$f_\ell(x) = (p_\ell)^x\, (1-p_\ell)^{1-x} \quad (\ell = A,B;\ x = 0,1;\ 0 \leq p_\ell \leq 1) \tag{2.16}$$

Note that in this example the parameter vector $\underline{\theta}$ is given by (p_A,p_B). If m learning objects per class are available of which n_A respectively n_B are one, the maximum likelihood estimates are given by

$$\hat{p}_\ell = \frac{n_\ell}{m} \quad (\ell = A,B) \tag{2.17}$$

The corresponding density estimates, using the plug-in rule, are

$$\hat{f}_\ell^{(1)}(x) = (\frac{n_\ell}{m})^x\, (1 - \frac{n_\ell}{m})^{1-x} \quad (\ell = A,B) \tag{2.18}$$

In order to use the estimates $\hat{f}_\ell^{(2)}(x)$ uniform a priori densities for p_A as well as for p_B will be assumed as an example.

$$h_\ell(p_\ell) = 1 \qquad 0 \leq p_\ell \leq 1 \qquad (\ell = A,B) \tag{2.19}$$

After some calculations using (2.7) and (2.5) can be found that

$$\hat{f}_\ell^{(2)}(x) = \left(\frac{n_\ell+1}{m+2}\right)^x \left(1 - \frac{n_\ell+1}{m+2}\right)^{1-x} \quad (\ell = A,B) \tag{2.20}$$

If the joint a priori density is also uniform, $h(p_A,p_B) = 1$, $0 \leq \{p_A,p_B\} \leq 1$, then (2.15) is valid and $\hat{f}_\ell^{(3)}(x) = \hat{f}_\ell^{(2)}(x)$, $\ell = A,B$. If $h(p_A,p_B)$ is uniform

k	m	$\tilde{\varepsilon}$ in %		
		estimators used		
		1	2	3
1	1	50.0	50.0	50.0
1	2	33.3	33.3	33.3
1	3	30.0	30.0	30.0
2	3	23.0	21.5	21.4
2	5	20.4	19.9	19.7
3	3	21.6	17.0	16.7

Table 2.1 Values of $\tilde{\varepsilon}$ (in %) for the presented example (see text) for feature size k and sample size m. The three estimators used are the ones defined by (2.3), (2.7) and (2.14).

along the line $p_A = 1-p_B$ and has a zero density elsewhere, then for the estimates is found

$$\hat{f}_A^{(3)}(x) = \left(\frac{m+n_A-n_B+1}{2m+2}\right)^x \left(1 - \frac{m+n_A-n_B+1}{2m+2}\right)^{1-x} \tag{2.21}$$

and

$$\hat{f}_B^{(3)}(x) = \left(\frac{m+n_B-n_A+1}{2m+2}\right)^x \left(1 - \frac{m+n_B-n_A+1}{2m+2}\right)^{1-x} \tag{2.22}$$

Notwithstanding the fact that the three estimates for $f_A(x)$ and $f_B(x)$ differ, they all yield the same discriminant function $\hat{S}(x)$ with

$$\begin{aligned} \hat{S}(1) &= C(n_A-n_B) \\ \hat{S}(0) &= C(n_B-n_A) \end{aligned} \tag{2.23}$$

where C is some positive constant. For the multivariate case, in which more than one feature is involved, the resulting discriminant functions are different. Some examples are presented in table 2.1 for the case of independent features. These values are exact computations of $\tilde{\varepsilon}$ for different feature sizes k and sample sizes m. The table shows that the results for the multivariate case differ for different estimators. More values are presented and discussed in chapter 4.

Chapter 3
SAMPLE SIZE

In this chapter the effect of the sample size on the classification error is considered using the density estimates described in the previous chapter. Especially the influence of the estimation errors on the density estimates will be considered.

In the first paragraph a general approach is given which results in a sample size dependent upper bound on the expected classification error $\bar{\varepsilon}$. In the next paragraph this is applied to a class of classification problems defined by Hughes [24]. The estimation accuracy for the case of normal distributions is considered in the third paragraph. Examples for nonparametric estimates are presented in paragraph 4. In the last paragraph some concluding remarks are made.

Parts of this chapter are already published in [12], [13] and [14].

3.1 A SAMPLE SIZE DEPENDENT ERROR BOUND

Error bounds are intensively studied in connection with feature extraction, see Fukunaga [23]. In that case known class distributions are assumed. These error bounds are therefore sample size independent. They do not take into account the errors made in estimating the distributions. For answering questions such as: What is the error caused by a finite learning set, or: What number of learning objects should be used in order to reach a certain accuracy, these error bounds are useless.

Effects of the sample size upon the accuracy obtained by an estimated discriminant function have been previously studied. Cover [9] gave for the total sample size (in our case 2m) a lower bound of two times the feature size. Otherwise learning sets of identically distributed classes become linearly separable, which is obviously absurd. Foley [22] presented

curves for the resubstitution error (the classification error estimated by classifying the learning objects) as a function of feature size and sample size. These curves are based on Monte Carlo experiments using identically distributed classes. From these curves can be concluded that a sample size feature size ratio of at least three or four is necessary. Extremely large figures are given by Hughes [24] and Abend et al. [1]. They compute what they call the optimal measurement complexity for a given sample size. The sample size for which a given measurement complexity is optimal appeared to be very large due to the extremely general model used.

In this paragraph an upper bound for the expected classification error is given. It is expressed into the Bayes error and the expected errors in the estimates of the class densities. The expected classification error, and thereby the upper bound is a function of the sample size. This makes it possible to compute the maximum number of learning objects necessary for a given value of the expected classification error.

The Bayes error can be written as given by (1.5).

$$\varepsilon^* = c \int_{S(\underline{x})<0} f_A(\underline{x})\, d\underline{x} + (1-c) \int_{S(\underline{x})\geq 0} f_B(\underline{x})\, d\underline{x} \tag{3.1}$$

The error made in the estimates $\hat{f}_A(\underline{x})$ and $\hat{f}_B(\underline{x})$ will be expressed into the Kolmogorov variational distance (see Fukunaga [23])

$$e_\ell = \tfrac{1}{2} \int_{\underline{X}} | \hat{f}_\ell(\underline{x}) - f_\ell(\underline{x}) |\, d\underline{x} \quad (\ell = A,B) \tag{3.2}$$

which is equivalent to

$$e_\ell = 1 - \int_{\underline{X}} \min \{\hat{f}_\ell(\underline{x}), f_\ell(\underline{x})\}\, d\underline{x} \quad (\ell = A,B) \tag{3.3}$$

This is called the *estimation error*. Note that e_ℓ is not a probability like ε. The definition of e_ℓ is such that $0 \leq e_\ell \leq 1$. In the case of perfect estimation e_ℓ is zero, for bad estimates e_ℓ approaches or is equal to one.

For the classification error caused by a discriminant function $\hat{S}(\underline{x})$ can be written (1.9).

$$\varepsilon = c \int_{\hat{S}(\underline{x})<0} f_A(\underline{x})\, d\underline{x} + (1-c) \int_{\hat{S}(\underline{x})\geq 0} f_B(\underline{x})\, d\underline{x} \tag{3.4}$$

This can be rewritten as

$$\varepsilon = c \int_{S(\underline{x})<0} f_A(\underline{x})\, d\underline{x} - c \int_{\substack{S(\underline{x})<0 \\ \hat{S}(\underline{x})\geq 0}} f_A(\underline{x})\, d\underline{x} + c \int_{\substack{S(\underline{x})\geq 0 \\ \hat{S}(\underline{x})<0}} f_A(\underline{x})\, d\underline{x} +$$

$$+ (1-c) \int_{S(\underline{x})\geq 0} f_B(\underline{x})\, d\underline{x} - (1-c) \int_{\substack{S(\underline{x})\geq 0 \\ \hat{S}(\underline{x})<0}} f_B(\underline{x})\, d\underline{x} + (1-c) \int_{\substack{S(\underline{x})<0 \\ \hat{S}(\underline{x})\geq 0}} f_B(\underline{x})\, d\underline{x} \tag{3.5}$$

Combination of some integrals and using (1.1) gives

$$\varepsilon = c \int_{S(\underline{x})<0} f_A(\underline{x})\, d\underline{x} + (1-c) \int_{S(\underline{x})\geq 0} f_B(\underline{x})\, d\underline{x} - \int_{\substack{S(\underline{x})<0 \\ \hat{S}(\underline{x})\geq 0}} S(\underline{x})\, d\underline{x} + \int_{\substack{S(\underline{x})\geq 0 \\ \hat{S}(\underline{x})<0}} S(\underline{x})\, d\underline{x} \tag{3.6}$$

The sum of the first two terms equals ε^*, see (1.5). If a region V is defined in which the classes are non-optimally classified by $\hat{S}(\underline{x})$,

$$V = \{\underline{x}: (S(\underline{x}) < 0 \wedge \hat{S}(\underline{x}) \geqslant 0) \vee (S(\underline{x}) \geqslant 0 \wedge \hat{S}(\underline{x}) < 0)\} \tag{3.7}$$

then (3.6) simplifies to

$$\varepsilon = \varepsilon^* + \int_V |S(\underline{x})|\, d\underline{x} \tag{3.8}$$

or

$$\varepsilon = \varepsilon^* + \int_V |c\, f_A(\underline{x}) - (1-c)\, f_B(\underline{x})|\, d\underline{x} \tag{3.9}$$

For $\underline{x} \in V$ the following inequality holds

$$|c\, \hat{f}_A(\underline{x}) - c\, f_A(\underline{x})| + |(1-c)\, \hat{f}_B(\underline{x}) - (1-c)\, f_B(\underline{x})| \geqslant$$

$$|c\, f_A(\underline{x}) - (1-c)\, f_B(\underline{x})| = |S(\underline{x})| \tag{3.10}$$

For the proof we distinguish two cases

a) $S(\underline{x}) < 0,\ \hat{S}(\underline{x}) \geqslant 0$ (3.11)

so

$$-S(\underline{x}) \leqslant \hat{S}(\underline{x}) - S(\underline{x}) \tag{3.12}$$

As both terms are positive, (3.12) is also true for the absolute values.

$$|S(\underline{x})| \leqslant |\hat{S}(\underline{x}) - S(\underline{x})| \tag{3.13}$$

or

$$|S(\underline{x})| \leqslant |c\ \hat{f}_A(\underline{x}) - (1-c)\ \hat{f}_B(\underline{x}) - c\ f_A(\underline{x}) + (1-c)\ f_B(\underline{x})| \tag{3.14}$$

$$\leqslant |c\ \hat{f}_A(\underline{x}) - c\ f_A(\underline{x})| + |(1-c)\ \hat{f}_B(\underline{x}) - (1-c)\ f_B(\underline{x})| \tag{3.15}$$

which proves (3.10)

b) $S(\underline{x}) \geqslant 0,\ \hat{S}(\underline{x}) < 0$ (3.16)

The proof is in this case similar to the one under a).

Substitution of (3.10) in (3.9) gives

$$\varepsilon \leqslant \varepsilon^* + \int_V \{|c\ \hat{f}_A(\underline{x}) - c\ f_A(\underline{x})| + |(1-c)\ \hat{f}_B(\underline{x}) - (1-c)\ f_B(\underline{x})|\}d\underline{x} \tag{3.17}$$

The integration area V can be equal to the whole space. An example is given in fig. 3.2. If V is replaced by the whole space one finds after using (3.2)

$$\varepsilon \leqslant \varepsilon^* + 2\ \{c\ e_A + (1-c)\ e_B\} \tag{3.18}$$

which completes the derivation of the upperbound on the classification error.

From (3.9) another upperbound can be found by immediately replacing V by the whole space,

$$\varepsilon \leqslant \varepsilon^* + \int_{\underline{x}} |c\ f_A(\underline{x}) - (1-c)\ f_B(\underline{x})|\, d\underline{x} \tag{3.19}$$

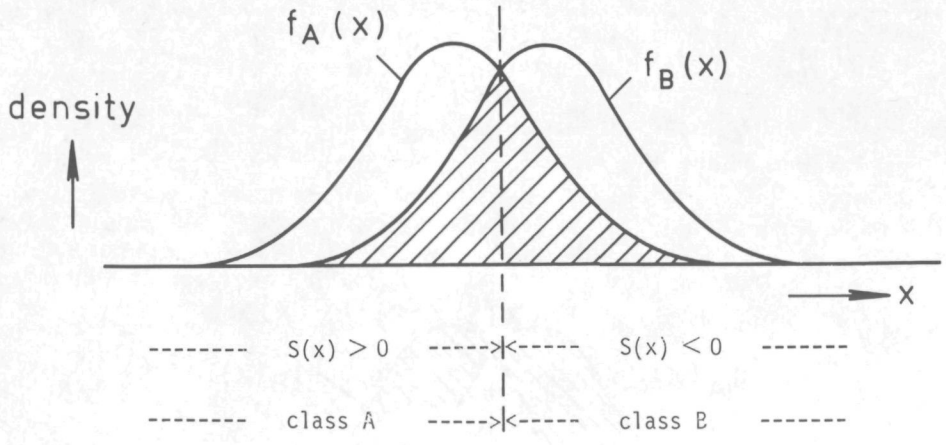


Fig. 3.1 The Bayes error ε^* equals half the shaded area ($c = 0.5$).

Substitution of (3.1) gives

$$\varepsilon \leq \int_{\underline{x}} \{\min \{c\ f_A(\underline{x}),\ (1-c)\ f_B(\underline{x})\} + |c\ f_A(\underline{x}) - (1-c)\ f_B(\underline{x})|\} d\underline{x} \tag{3.20}$$

This is equivalent to

$$\varepsilon \leq \int_{\underline{x}} \max \{c\ f_A(\underline{x}),\ (1-c)\ f_B(\underline{x})\} d\underline{x} \tag{3.21}$$

or

$$\varepsilon \leq \int_{\underline{x}} \{c\ f_A(\underline{x}) + (1-c)\ f_B(\underline{x}) - \min \{c\ f_A(\underline{x}),\ (1-c)\ f_B(\underline{x})\}\}\ d\underline{x} \tag{3.22}$$

or

$$\varepsilon \leq 1 - \varepsilon^* \tag{3.23}$$

The upper bound is reached when the classification regions corresponding with ε^* are completely reversed, see figs. 3.1 and 3.2. Together with the obvious fact that $\varepsilon^* \leq \varepsilon$ one gets from (3.23) and (3.18)

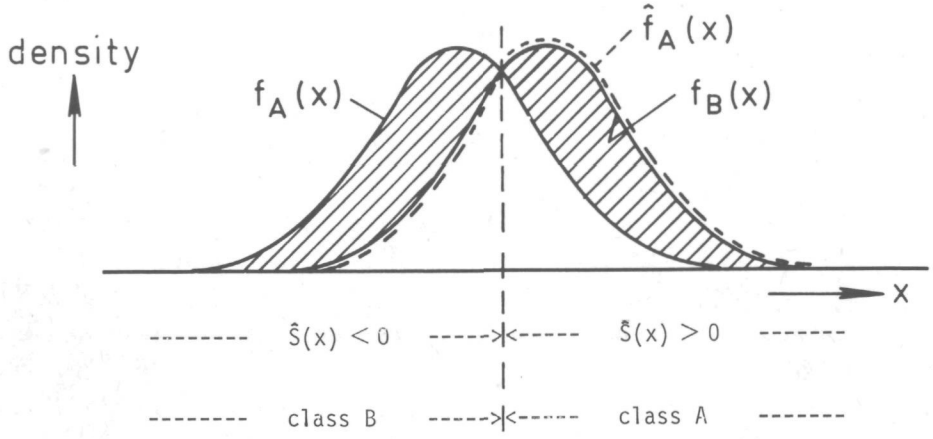


Fig. 3.2 Example of the result of an extremely bad estimate for one of the classes. The interrupted line shows the estimated density for class A. The density function of class B is assumed to be estimated perfectly. So e_A equals half the shaded area and $e_B = 0$. Comparison with fig. 3.1 shows ($c = 0.5$)

1) $\varepsilon = 1 - \varepsilon^*$

2) $\varepsilon = \varepsilon^* + e_A + e_B$

By this the two upper bounds (3.23) and (3.18) are reached simultaneously.

$$\varepsilon^* \leq \varepsilon \leq \min \{(1-\varepsilon^*),\ \varepsilon^* + 2(c\ e_A + (1-c)e_B)\} \tag{3.24}$$

For most practical problems, however (3.18) is a more stringent bound than (3.23). In fig. 3.2 an example is given where the upper bound (3.18) is reached. In this example the densities and their estimates of the two equally probable classes A and B are given. The density estimate of class B is equal to the true density. The density estimate of class A is such that the classification compared with the one based on the true densities is reversed. From the figure can be understood that by this the additional classification error $\varepsilon - \varepsilon^*$ equals e_A. As $e_B = 0$ the upper bound $\varepsilon = \varepsilon^* + e_A + e_B$ is reached. Note that in this example also $\varepsilon = 1 - \varepsilon^*$, so the two upper bounds coincide.

A bound on the classification error expressed in e_A and e_B has hardly any practical value, because e_A and e_B can, in general, have any value between zero and one and are in a particular problem unknown. A bound on the expected error would be more useful. It gives an indication of the expected accuracy. Such a bound can easily be found by taking the expectation of (3.18).

$$\bar{\varepsilon} = E_X(\varepsilon) \leq \varepsilon^* + 2\ \{c\ E_X(e_A) + (1-c)\ E_X(e_B)\} \tag{3.25}$$

This bound will not be very tight because $\hat{f}(\underline{x})$ can deviate from $f(\underline{x})$ to two sides. Only one of these sides can cause an erroneous classification.
An exact expression for $\bar{\varepsilon}$ can be found by taking the expectation of (3.8) over the learning set. Therefore (3.6) is rewritten as

$$\varepsilon = \varepsilon^* + \int_{\substack{S(\underline{x})\geq 0 \\ \hat{S}(\underline{x})<0}} S(\underline{x})\ d\underline{x} - \int_{\substack{S(\underline{x})<0 \\ \hat{S}(\underline{x})\geq 0}} S(\underline{x})\ d\underline{x}$$

The expectation over the learning set effects only $S(\underline{x})$, of which only its sign is relevant.
So

$$\bar{\varepsilon} = \varepsilon^* + \int_{S(\underline{x})\geq 0} \text{Prob}\ (\hat{S}(\underline{x}) < 0)\ S(\underline{x})d\underline{x} \ +$$

$$- \int_{S(\underline{x})<0} \text{Prob}\ (\hat{S}(\underline{x}) \geq 0)\ S(\underline{x})d\underline{x} \tag{3.26}$$

If one assumes that the probabilities in (3.26) can be 0.5 at most, which is for the given integration areas likely but not necessarily true, $\bar{\varepsilon}$ is bounded by

$$\bar{\varepsilon} \leq \varepsilon^* + 0.5 \int_{\underline{x}} |S(\underline{x})| \, d\underline{x}$$

Using (3.10) one obtains

$$\bar{\varepsilon} \leq \varepsilon^* + c \, E_\chi(e_A) + (1-c) \, E_\chi(e_B) \tag{3.27}$$

which is a factor two better than (3.25).

The next paragraphs will investigate for some special cases, how $E_\chi(e_A)$ and $E_\chi(e_B)$ depend on sample size and feature size. This results into sample sizes which guarantee, in expectation, a certain classification accuracy.

3.2 THE ERROR BOUND FOR A GENERAL MEASUREMENT SPACE

We will adopt here a model originally presented by Hughes [24]. Let x be a measurement outcome into one of n cells with probabilities p_A^j and p_B^j (j = 1,n) for the classes A and B. n is called the *measurement complexity*. It can be compared with the dimensionality k of, for instance, continuous spaces. The influence of both on the number of distributional parameters is similar. Assume that 2m objects are available for the estimation of p_A^j and p_B^j. Maximum likelihood estimates indicated by $\hat{p}_A^j$ and $\hat{p}_B^j$ will be used. For the estimation error of class ℓ (ℓ = A,B) can be written, (compare (3.3))

$$e_\ell = 1 - \sum_{j=1}^{n} \min \{\hat{p}_\ell^j, p_\ell^j\} \tag{3.28}$$

By taking the expectation over all learning sets one gets

$$E_\chi(e_\ell) = 1 - \sum_{j=1}^{n} E_\chi \{ \min\{\hat{p}_\ell^j, p_\ell^j\}\} \tag{3.29}$$

Define

$$y = (\hat{p}_\ell^j - p_\ell^j) \{p_\ell^j(1-p_\ell^j)/m\}^{-\frac{1}{2}} \tag{3.30}$$

So

$$E_\chi\{\min\{\hat{p}_\ell^j, p_\ell^j\}\} = E\{\min\{y,0\}\}\ \{p_\ell^j(1-p_\ell^j)/m\}^{\frac{1}{2}} + p_\ell^j \tag{3.31}$$

If m is large enough $\hat{p}_\ell^j$ is approximately normally distributed with expectation p_ℓ^j and variance $p_\ell^j(1-p_\ell^j)/m$. In that case y has approximately a standard normal distribution. For the expectation of min{y,0} is found

$$E\{\min\{y,0\}\} = \int_{-\infty}^{0} (2\pi)^{-\frac{1}{2}}\ y\ \exp(-y^2/2)dy = -(2\pi)^{-\frac{1}{2}} \tag{3.32}$$

Using (3.31) and (3.32), (3.29) becomes

$$E_\chi(e_\ell) = \sum_{j=1}^{n} \{p_\ell^j(1-p_\ell^j)/(2\pi m)\}^{\frac{1}{2}} \tag{3.33}$$

because

$$\sum_{j=1}^{n} p_\ell^j = 1$$

In appendix A it is shown that, for $c = \frac{1}{2}$, $c\ E_\chi(e_A) + (1-c)\ E_\chi(e_B)$ is maximum if for n/2 values of j (n even)

$$p_A^j = 2\varepsilon^*/n$$

and (3.34)

$$p_B^j = 2(1-\varepsilon^*)/n$$

and if for the other n/2 values of j

$$p_A^j = 2(1-\varepsilon^*)/n$$

and (3.35)

$$p_B^j = 2\varepsilon^*/n$$

Using (3.33) - (3.35) for (3.25) can be written

$$\bar{\varepsilon} \leq \varepsilon^* + n(2\pi m)^{-\frac{1}{2}}\{\{2\varepsilon^*(1-2\varepsilon^*/n)/n\}^{\frac{1}{2}} + \{2(1-\varepsilon^*)(1-2(1-\varepsilon^*)/n)/n\}^{\frac{1}{2}}\} \tag{3.36}$$

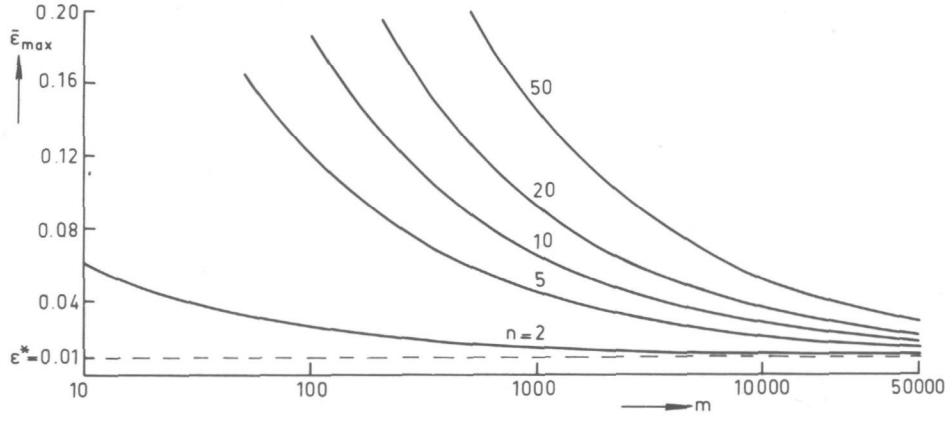


Fig. 3.3.a

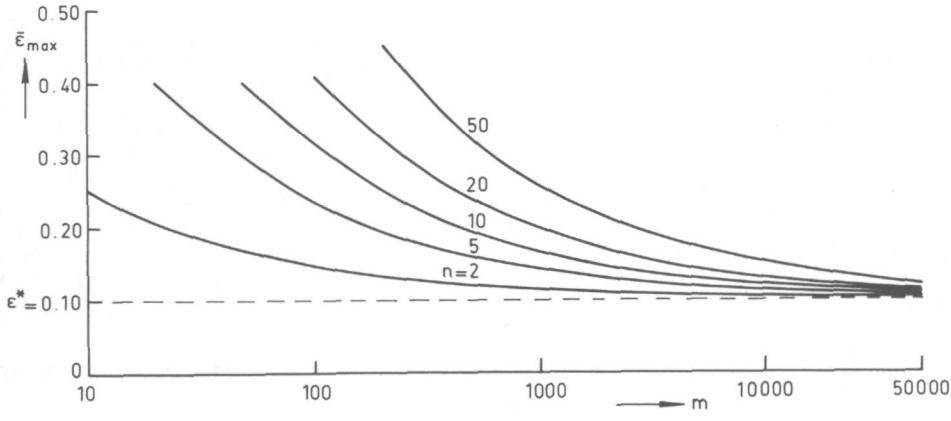


Fig. 3.3.b

Fig. 3.3 The upper bound of $\bar{\varepsilon}$ for the general measurement space (3.36) as a function of sample size m and measurement complexity n.

a. $\varepsilon^* = 0.01$

b. $\varepsilon^* = 0.1$.

For large values of n this reduces to

$$\bar{\varepsilon} \leqslant \varepsilon^* + \{n/(\pi m)\}^{\frac{1}{2}} \{(\varepsilon^*)^{\frac{1}{2}} + (1-\varepsilon^*)^{\frac{1}{2}}\} \qquad (3.37)$$

In fig. 3.3 the upper bound (3.36) of $\bar{\varepsilon}$ is given for two values of ε^* as a function of sample size m and measurement complexity n. These curves should be interpreted in the following way. If the Bayes error ε^* (infinite sample size case) equals 0.1 then fig. 3.3b gives the sample size that guarantees an inaccuracy of less than $\bar{\varepsilon}_{max}$ for measurement complexity n. Because the Bayes error is only rarely known in practice, the given curves serve to give an impression of the sample size needed for a certain accuracy by given measurement complexity n. For small sample sizes the approximation of y by a normal distribution, and thereby the expressions (3.32) and (3.36), become inaccurate.

The resulting numbers of learning samples are large and in many practical problems not available. Additionally, they are with respect to many practical results extremely pessimistic. This is caused by the very general model which covers many difficult classification problems and by the worst case approach that has been followed.

In order to illustrate this last statement the exact value of $\bar{\varepsilon}$ given by (3.26) has been computed approximately for the same probability distribution, defined by (3.34) and (3.35), as used above. The results, shown in fig. 3.4 have to be compared with those of fig. 3.3, where the upper bound is given, which is based on the same probability distribution. Formula (3.26) has been approximated for this case using the Camp-Paulson approximation for the cumulative binomial distribution as given by Molenaar [29].

The resulting values of $\bar{\varepsilon}$ for given m, n and ε^*, as follow from fig. 3.4, are much closer to values obtained for ε in practical situations than the ones of fig. 3.3. This illustrates how pessimistic the upper bound is.

The results of the upper bound are based on a worst case approach and have to be valued in that light. The results actually obtained in a particular classification problem are, with high probability, much better. The upper bound, however, gives a guarantee for the resulting values of $\bar{\varepsilon}$ as a function of m and n.

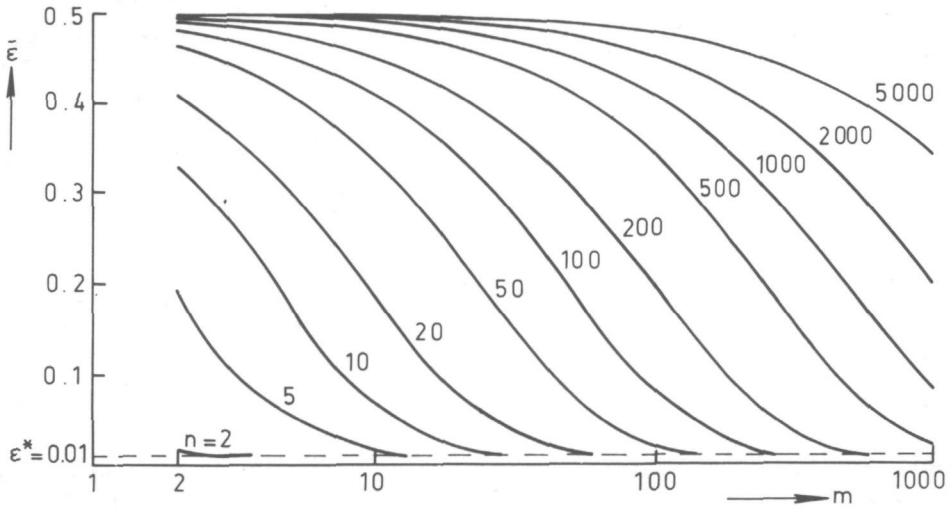


Fig. 3.4.a

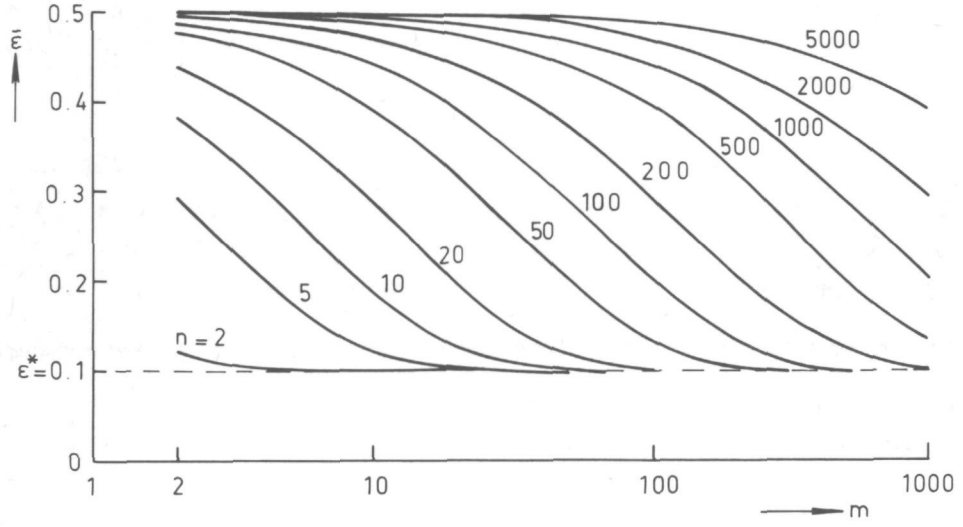


Fig. 3.4.b

Fig. 3.4 The expected classification error $\bar{\varepsilon}$ for the function defined by (3.34) and (3.35) as a function of sample size m and measurement complexity n

a. $\varepsilon^* = 0.01$

b. $\varepsilon^* = 0.1$.

3.3 THE CLASSIFICATION ERROR USING NORMALLY DISTRIBUTED FEATURES

The expectation of the estimation error e for an arbitrary k-dimensional normal distribution can be written, as follows from (3.3), as

$$E_\chi(e) = 1-E_\chi\left\{ \int_{\underline{x}} \min\{f(\underline{x}|\hat{\underline{\mu}},\hat{\Sigma}), f(\underline{x}|\underline{\mu},\Sigma)\}d\underline{x} \right\} \tag{3.38}$$

in which $f(\underline{x}|\underline{\mu},\Sigma)$ is the normal density function with expectation $\underline{\mu}$ and covariance matrix Σ. In this paragraph the plug-in rule based on the maximum likelihood estimates $\hat{\underline{\mu}}$ and $\hat{\Sigma}$ will be used only.

In appendix D it is shown that (3.38) is independent of μ and Σ. Therefore a multidimensional standard normal distribution may be chosen for $f(\cdot)$. E_χ can be written in that case as

$$E_\chi(e) = 1-E_\chi\left\{ \int_{\underline{x}} \min\{f(\underline{x}|\hat{\underline{\mu}},\hat{\Sigma}),\ f(\underline{x}|\underline{0},I)\}d\underline{x} \right\} \tag{3.39}$$

in which I is the identity matrix. Because of this $E_\chi(e)$ only depends on the dimensionality k and the sample size m.

We computed $E_\chi(e)$ as a function of m and k using Monte Carlo procedures. The integral of the minimum in (3.39) was approximated by using 2×50 randomly selected points according $f(\underline{x}|\underline{0},I)$ and its estimate. This procedure is explained in appendix E. The expectation was obtained by averaging the results of 200 randomly chosen learning sets of size m. The accuracy of this method can be found by computing the standard deviation of those 200 results. In fig. 3.5 $E_\chi(e)$, estimated in this way, is shown as a function of m and k (see also table 3.1). The values of m can,for our purposes, be interpreted as that number of learning objects which guarantees that in expectation the contribution of the estimation error of some normal density function to the expected classification error is less than 2 Prob ($\underline{x} \in$ class ℓ) $E_\chi(e_\ell)$, see (3.25).

In 3.1 it was stated that the upper bound (3.25) will probably be too loose. In order to get some impression of this the following experiments were performed for the case of normal distributions. For a number of classification problems with $c=\frac{1}{2}$ and randomly chosen learning sets, ε^*, ε, e_A, e_B and γ, defined by

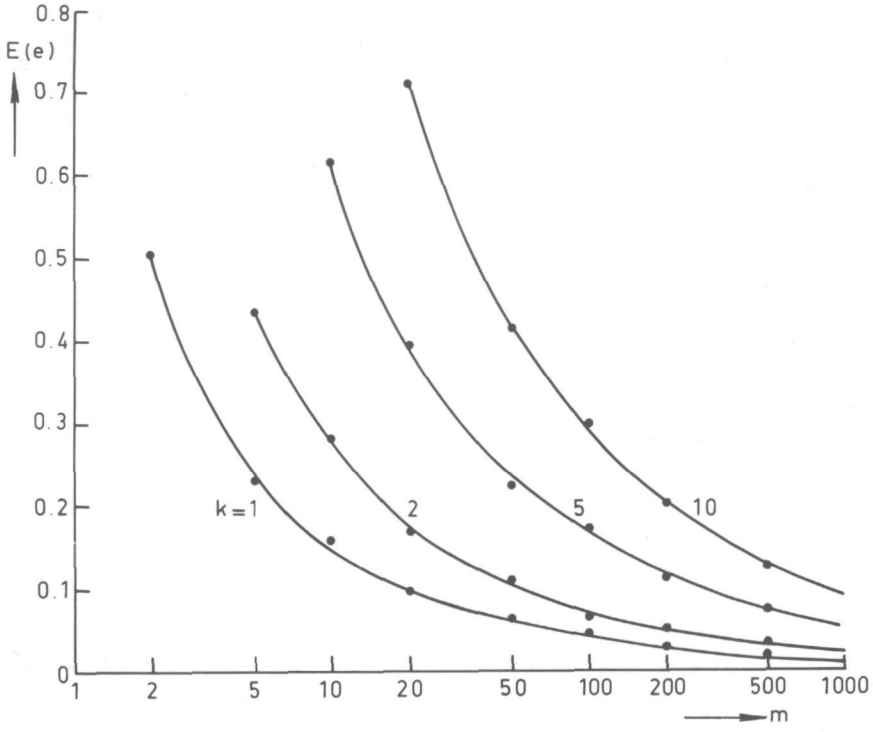


Fig. 3.5 The expected estimation error $E_\chi(e)$ for normal distributions with dimensionality k and sample size m.

m	k			
	1	2	5	10
2	0.506 (0.019)			
5	0.233 (0.011)	0.437 (0.013)		
10	0.160 (0.008)	0.284 (0.009)	0.617 (0.009)	
20	0.098 (0.008)	0.170 (0.008)	0.396 (0.008)	0.712 (0.006)
50	0.065 (0.007)	0.111 (0.007)	0.226 (0.008)	0.416 (0.007)
100	0.047 (0.007)	0.068 (0.007)	0.173 (0.007)	0.289 (0.007)
200	0.031 (0.007)	0.052 (0.007)	0.114 (0.007)	0.203 (0.007)
500	0.019 (0.007)	0.034 (0.007)	0.074 (0.007)	0.126 (0.007)
1000	0.009 (0.007)	0.021 (0.007)	0.051 (0.007)	0.097 (0.007)

Table 3.1 The expected estimation error $E_\chi(e)$ for normal distributions with dimensionality k and sample size m. The presented values are the mean results of a Monte Carlo simulation. Between the brackets the computed standard deviations of the means are given.

			m = 20			m = 50		
μ	ω	ε^*	γ 1)	2)	3)	γ 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 3.2 Results of a number of two dimensional experiments, each repeated for ten different learning sets and for sample sizes of 20 and 50. The distributions are normal and independent with means (0,0) and (μ,0) and with variances (1,1) and (ω,1).
1) mean value of γ in ten experiments
2) number of times $\gamma > 0.15$
3) number of times $\gamma > 0.20$

$$\varepsilon = \varepsilon^* + \gamma(e_A+e_B) \tag{3.40}$$

were computed. The resulting values of γ appeared very often to be less than 0.2. In table 3.2 the results of an example are presented where the distribution of the classes A and B are both binormal with zero correlation. A has mean (0,0) and variances (1,1) and B had mean (μ,0) and variances (ω,1). For each value of μ and ω ten learning sets were chosen at random, resulting in ten values of γ. In table 3.2 the mean value of γ and the number of times that γ was larger than 0.15 or 0.20 are given. The results are presented for sample sizes of 20 and 50. These experiments show that under certain conditions the accuracy is much greater than can be determined from fig. 3.5 and formula (3.25).

In literature much attention has been paid to the behaviour of ε as a function of m, k and ε^* for the case of normally distributed classes with equal covariance matrices. A relation to the estimation error, however, was not found, yet. An asymptotic expansion of $\bar{\varepsilon}$, up to the second order with respect to m^{-1}, has been given by Okamoto [31]. The asymptotic distribution of ε has been studied by Lachlan [27]. Monte Carlo experiments for several values of ε^*,m

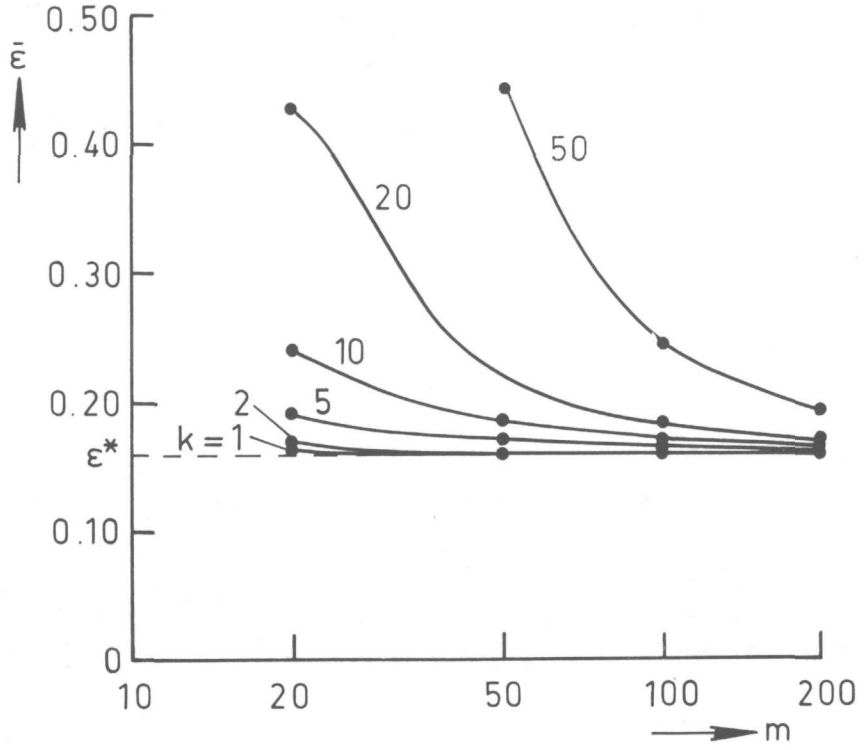


Fig. 3.6 The expected classification error $\bar{\varepsilon}$ for two normal distributions with equal covariance matrices I and with means on a distance of two, as a function of feature size k and sample size m.

k	m			
	20	50	100	200
1	0.161 (0.003)	0.160 (0.001)	0.159 (0.001)	0.159 (0.001)
2	0.168 (0.009)	0.162 (0.004)	0.161 (0.003)	0.159 (0.001)
5	0.190 (0.019)	0.170 (0.008)	0.165 (0.005)	0.161 (0.002)
10	0.244 (0.047)	0.185 (0.014)	0.171 (0.006)	0.165 (0.003)
20	0.430 (0.117)	0.219 (0.023)	0.186 (0.009)	0.171 (0.005)
50		0.445 (0.067)	0.247 (0.022)	0.194 (0.007)

Table 3.3 The estimation and its standard deviation of the expected classification error $\bar{\varepsilon}$, for two normal distributions with equal covariance matrices I and with means on a distance of two, as a function of feature size k and sample size m.

and k are presented by Dunn [17], Bouillon et. al [4], and in a wider context by Van Ness and Simpson [44]. The performance of a number of linear discriminant functions in the normally distributed case are compared by Sorum [38]. However, all these references do not enable us to present $\bar{\varepsilon}$ as a function of sample size and feature size on the basis of published results. This is due to the accidental choices made for m and k. We were, therefore, forced to run our own Monte Carlo experiments. They are based on two equally probable classes, both normally distributed, with the identity matrix as covariance matrix and with means on a distance of one. 50 different learning sets of size m were generated and each time ε was computed analytically on the basis of the estimated means and covariance matrix, in the same way as published by Dunn [17]. The results are shown in fig. 3.6 and in table 3.3. Comparison with the results obtained by substitution of the data of fig. 3.5 into the upper bound (3.25) shows again a wide gap between the bound and the actually obtained results.

Finally we will give a short comment on the procedure followed by Bouillon et al. [4]. In contrast with the earlier paper by Dunn [17] they find an estimate of $\bar{\varepsilon}$ by averaging the classification errors made in only one of the two classes. This error has a range of 0 to 1. It is therefore possible that their estimate of $\bar{\varepsilon}$ is lower than ε^*. In fact this happens several times in their published results, but is not commented on by the authors. If they had averaged the values of ε, which have a range of ε^* to 1, as is done here and by Dunn [17], more realistic figures would have been obtained.

3.4 THE CLASSIFICATION ERROR USING NONPARAMETRIC ESTIMATES

In this paragraph some considerations will be given to the accuracy of density estimates by using histograms and Parzen estimators. This will be illustrated by some experimental results using Monte Carlo procedures on normal densities.

The histogram as a density estimator shows similar characteristics as the general measurement space described in 3.2. In both cases the number of learning objects in a cell is used for estimating the probability of finding an object in that cell. The estimates converge in both cases in the same way to those probabilities. The difference is, however, that the general measurement space is intrinsically discrete while the histogram may be an approximation of a continuous stochastic variable. The result is that, for a constant number

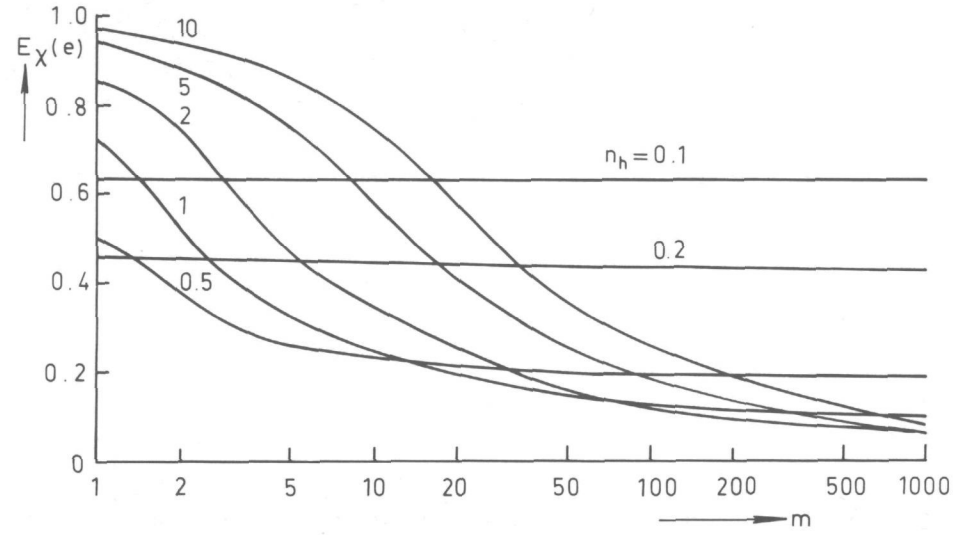


Fig. 3.7 The expected estimation error of a histogram of a normal distribution for a number of values of n_h (the number of cells used per standard deviation), m is the sample size.

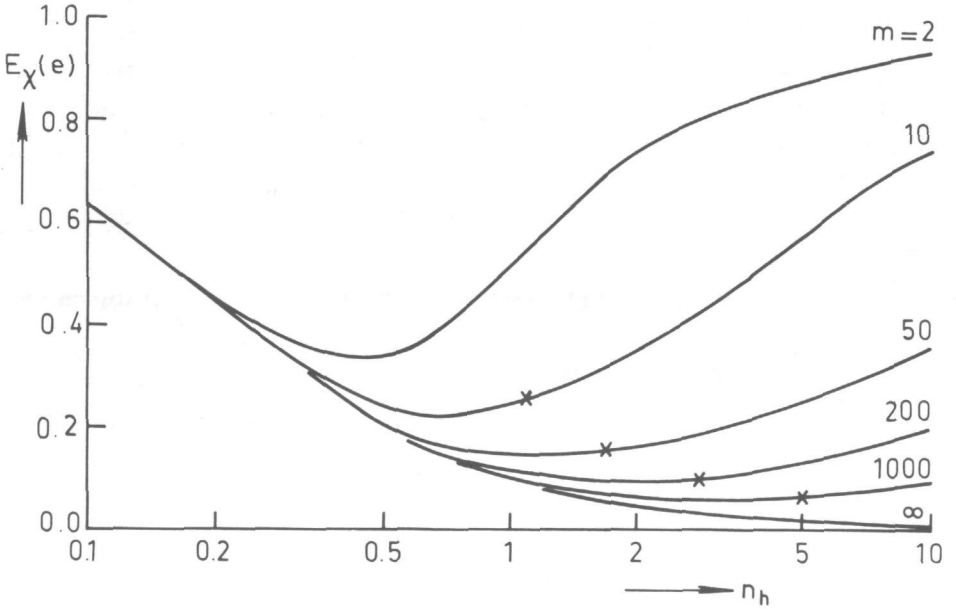


Fig. 3.8 The expected estimation error of a histogram of a normal distribution for a number of values of the sample size m, n_h is the number of cells used per standard deviation. The indicated points correspond with a interval width of the expectation of $(x_{max}-x_{min})/\sqrt{m}$.

of histogram cells and a constant width of those cells, the histogram cannot represent the density exactly, even for very large numbers of objects. Always some residual error remains to exist. This error approaches zero only if by increasing number of objects the number of cells approaches infinity and the width of those cells approaches zero. The remarks and figures of 3.2 apply, because of the above, for the histogram approach except for the residual error.

The estimation error of a one dimensional histogram for an underlying normal distribution with standard deviation σ has been calculated using a Monte Carlo procedure. In this experiment n_h cells were chosen on a length of one standard deviation. So the width of a histogram cell is σ/n_h. The location of the mean of the distribution was random in relation to the cells. For each value of n_h a sample set of size m was generated 50 times. The estimation errors, calculated by integration, were averaged. The results are shown in fig. 3.7 as a function of m for a number of values of n_h. The convergence of $E_\chi(e)$ can be studied using this figure. An optimal choice for n_h is best made from fig.3.8 where the expected estimation error is given as a function of n_h for a number of values of m. In a practical situation with unknown distributions the optimal number of cells has to be chosen using a priori knowledge, or from previous experiments.

As a rule of thumb sometimes $\Delta x = (x_{max} - x_{min})/\sqrt{m}$ is used for the interval width of a histogram, in which x_{max} and x_{min} are the maximum and minimum values in the sample set. This corresponds with the indicated points in fig. 3.8. These points are close to the optimal ones.

An example of the application of a histogram for discriminant analysis is given by Moss [30], who maps continuous signals into a discrete space. He experimentally investigates the influence of the number of cells on the error. In such a discriminant analysis, cells in the non-overlapping regions of the class distributions can be combined for efficient coding, because this will not influence the discriminant error. This causes a lower optimum number of cells in the discriminant problem compared with the density estimation problem.

In the case of density estimates using *Parzen estimators* the situation is slightly different. Such an estimate can be written as (see Fukunaga [23])

$$\hat{f}(\underline{x}|h,\chi) = \frac{1}{m} \sum_{i=1}^{m} u(\underline{x}-\underline{x}^i|h) \qquad (3.41)$$

Several choices can be made for the so called kernel function $u(\underline{x}|h)$ in which h is a width parameter, often called the smoothing parameter. A common choice for $u(\cdot)$ is the normal density function with mean zero and covariance matrix

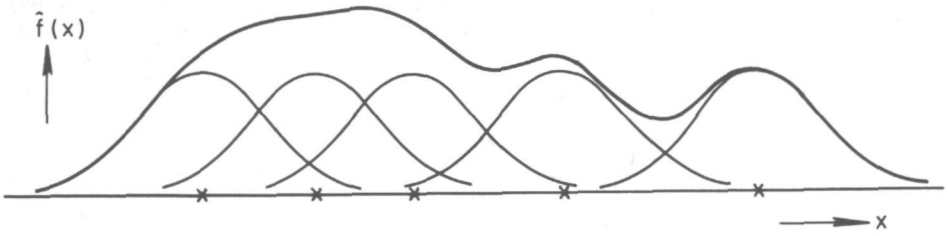


Fig. 3.9 A one dimensional example of a Parzen estimation of a density function using five objects.

h^2I, when I is the identy matrix.† A one dimensional example is given in fig. 3.9. A problem is the choice of h. In [13] a pseudo maximum likelihood method has been proposed, which will be used below in one of the experiments.

In order to be able to compare the estimation errors of a histogram and a Parzen estimate the following one dimensional experiment was performed. For a number of values of h a sample set of size m was generated 50 times, using a normal distribution. The estimation errors, determined by a Monte Carlo procedure (see appendix E) using 50 samples for the true density function and 50 samples for its estimate, were averaged. The results are shown in fig. 3.10 for a normal kernel and fig. 3.11 for a uniform kernel. Comparison with fig. 3.8 learns that the results of the Parzen estimates are better than the results of the histogram estimates.

The figures 3.10 and 3.11 show that except for very small values of 1/h approximately the same accuracy is reached for the two types of kernels. The normal kernels, however, give a somewhat lower minimum error. This is caused by the fact that the density function is estimated by kernels of the same shape (both normal). This benefit of the normal kernels, however, is rather small.

The performance of the Parzen estimator on a multidimensional normal density using a normal kernel and using the pseudo maximum likelihood estimator for h is shown in fig. 3.12. These results are found by generating a sample set of size m 50 times. Each time a new estimation of h is made. The estimation

---

† In our experiments the kernel function was always chosen to be such that $u(\underline{x}|h) = \prod_{j=1}^{k} u'(x_j|h)$, in which $u'(\cdot)$ is a one-dimensional density function with standard deviation h.

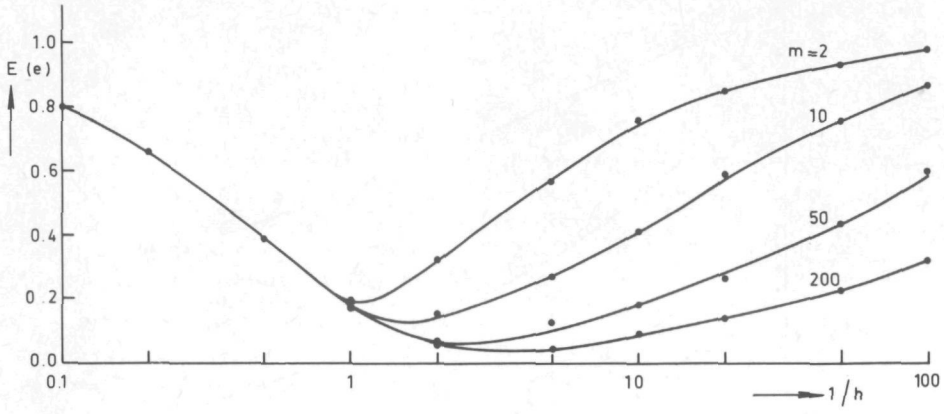


Fig. 3.10 The expected estimation error of a Parzen estimation of a one dimensional normal distribution as a function of the inverse of the smoothing parameter for a number of sample sizes m. Normal kernels are used.

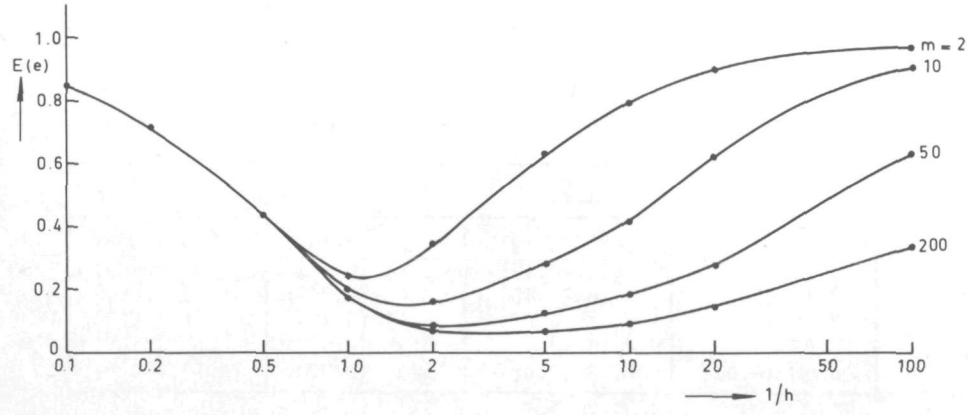


Fig. 3.11 The expected estimation error of a Parzen estimation of a one dimensional normal distribution as a function of the inverse of the smoothing parameter for a number of sample sizes m. Uniform kernels are used.

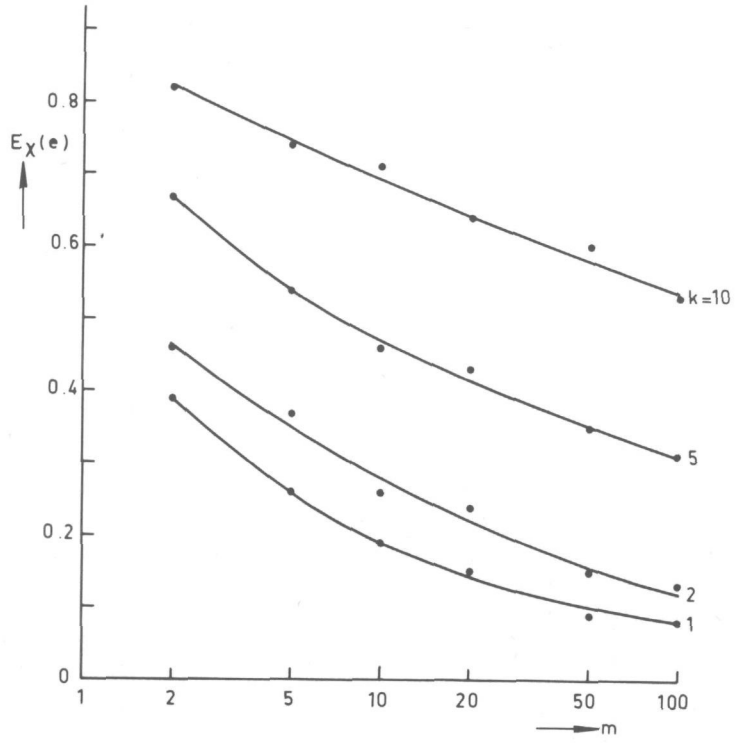


Fig. 3.12 The expected estimation error of a Parzen estimation of a k-dimensional normal distribution as a function of the sample size m. Normal kernels are used. The smoothing parameter is estimated using the method described in [13].

m	k			
	1	2	5	10
2	0.39 (0.20)	0.46 (0.19)	0.67 (0.13)	0.82 (0.08)
5	0.26 (0.10)	0.37 (0.13)	0.54 (0.11)	0.74 (0.09)
10	0.18 (0.10)	0.26 (0.10)	0.46 (0.09)	0.71 (0.07)
20	0.15 (0.09)	0.24 (0.09)	0.43 (0.09)	0.64 (0.07)
50	0.09 (0.07)	0.15 (0.07)	0.35 (0.07)	0.60 (0.05)
100	0.08 (0.07)	0.13 (0.07)	0.31 (0.07)	0.53 (0.06)

Table 3.4 The expected estimation error $E_\chi(e)$ of a Parzen estimation of a k-dimensional normal distribution using a learning set of size m. Normal kernels are used. The smoothing parameter is estimated using the method described in [13]. The presented values are the mean results of a Monte Carlo simulation. Between the brackets the computed standard deviations of the means are given.

errors are computed by Monte Carlo procedures using (see appendix E) 50 samples for the true density and 50 samples for its estimate. The results are averaged over the 50 runs. They are also shown in table 3.4. They have to be compared with the results of the parametric estimate shown in fig. 3.5. For small sample sizes these results are comparable or better due to the fact that the chosen kernel and the true density function are identical. The parametric estimates converge, of course, much faster than the Parzen estimates. These experiments use much computing time and are, therefore, not run for sample sizes larger than 100 and not repeated more than 50 times. We realize, however, that by this the obtained accuracy is not very high.

In order to illustrate the relative value of a good density estimate for classification the following classification experiments are performed. From two five dimensional normal densities, each with the unity matrix as covariance matrix and with their means on a distance of two, 2×50 learning objects are generated. For a number of values of h, Parzen estimates are computed for the case of normal kernels as well as for the case of uniform kernels. The resulting discriminant function given by (1.7) and (3.41), in which $c = 0.5$ is chosen, is tested by 2×1000 test objects. This is repeated ten times for ten different learning sets. Always the same test set is used. The averaged classification results, which are an estimate for $\bar{\varepsilon}$, are shown in fig. 3.13 for the normal kernel and in fig. 3.14 for the uniform kernel (see also table 3.5 and table 3.6). The strong difference between the two results is explained below.

The estimation errors for one of the two classes are computed by a Monte Carlo procedure using 50 objects for the true density estimate and 50 objects for its estimate (see appendix E). The averaged results over the ten experiments are shown in fig. 3.15 for the normal kernel and in fig. 3.16 for the uniform kernel (see also table 3.7 and table 3.8). There is little difference between the two curves, which indicates that, at least for 50 objects in a five dimensional space, the choice of the shape of the kernel is not very important for the accuracy of the density estimation.

The difference between the results for the estimation error and the classification error can be understood from studying the kernel properties. By increasing smoothing parameter the normal kernel is better and better approximated, in the area of interest, by the linear term in its Taylor expansion. The discriminant function approaches, therefore, the perpendicular bisector between the two means (see Specht [39]). This happens to be, in the case of the presented example, the optimal discriminant function. In the case

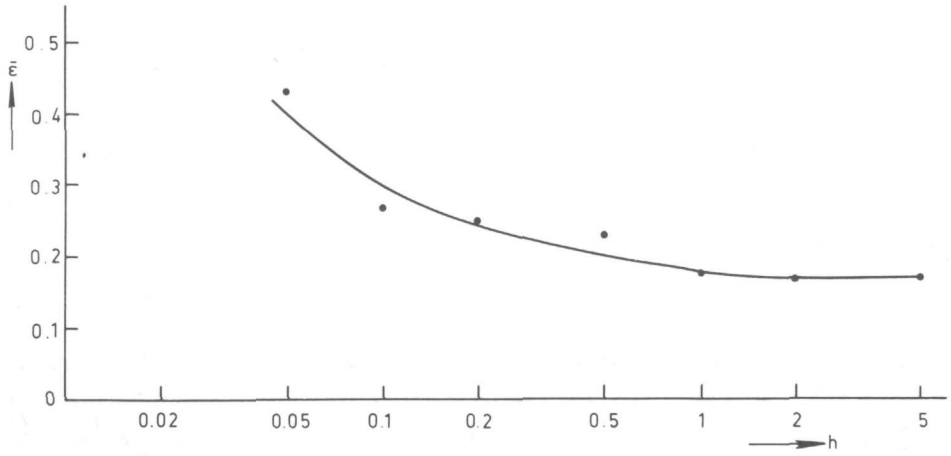


Fig. 3.13 The expected classification error in a five dimensional example as a function of the smoothing parameter of the Parzen estimation. Normal kernels are used. m = 50.

h	$\bar{\varepsilon}$
0.05	0.43 (0.01)
0.10	0.27 (0.02)
0.20	0.25 (0.02)
0.50	0.23 (0.02)
1.00	0.18 (0.01)
2.00	0.17 (0.01)
5.00	0.17 (0.01)

Table 3.5 The expected classification error $\bar{\varepsilon}$ in a five dimensional example as a function of the smoothing parameter of the Parzen estimation (m = 50). Normal kernels were used. The presented values are the mean results of a Monte Carlo simulation. Between the brackets the computed standard deviations of the means are given.

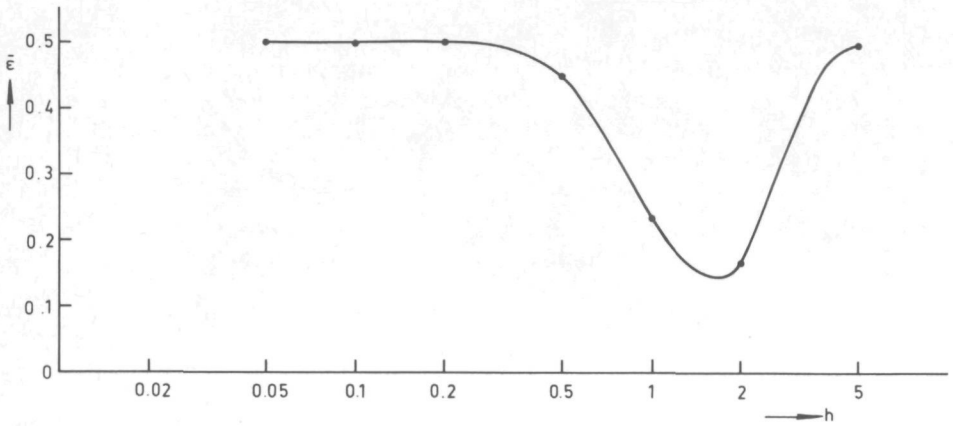


Fig. 3.14 The expected classification error in a five dimensional example as a function of the smoothing parameter of the Parzen estimation. Uniform kernels are used. m = 50.

h	$\bar{\varepsilon}$
0.05	0.500 (0.000)
0.10	0.500 (0.000)
0.20	0.500 (0.001)
0.50	0.452 (0.011)
1.00	0.234 (0.016)
2.00	0.169 (0.011)
5.00	0.500 (0.001)

Table 3.6 The expected classification error $\bar{\varepsilon}$ in a five dimensional example as a function of the smoothing parameter of the Parzen estimation (m = 50). Uniform kernels were used. The presented values are the mean results of a Monte Carlo simulation. Between the brackets the computed standard deviations of the means are given.

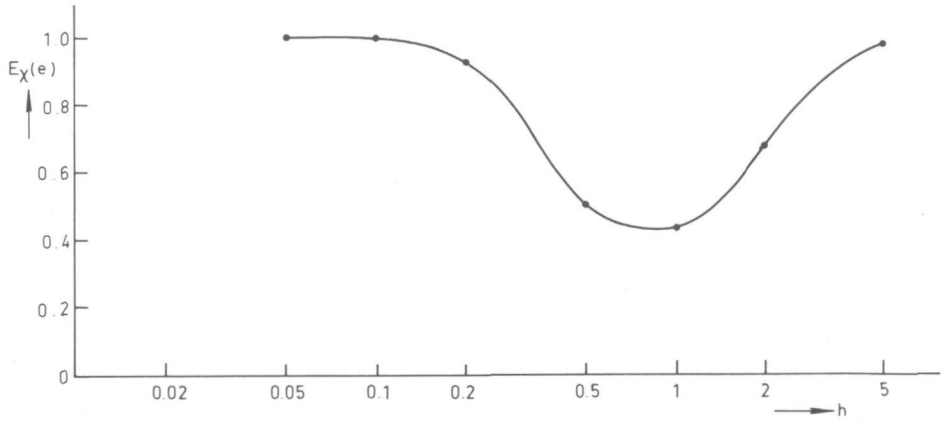


Fig. 3.15 *The expected estimation error of one of the classes in a five dimensional classification problem as a function of the smoothing parameter of the Parzen estimation. Normal kernels are used. m = 50.*

h	$E_\chi(e)$
0.05	1.00 (0.00)
0.10	1.00 (0.00)
0.20	0.93 (0.01)
0.50	0.49 (0.02)
1.00	0.44 (0.02)
2.00	0.78 (0.03)
5.00	0.98 (0.01)

Table 3.7 *The expected estimation error $E_\chi(e)$ of a five dimensional normal distribution using a Parzen estimation with normal kernels and smoothing parameter h (m = 50). The presented values are the mean results of a Monte Carlo simulation. Between the brackets the computed standard deviations of the means are given.*

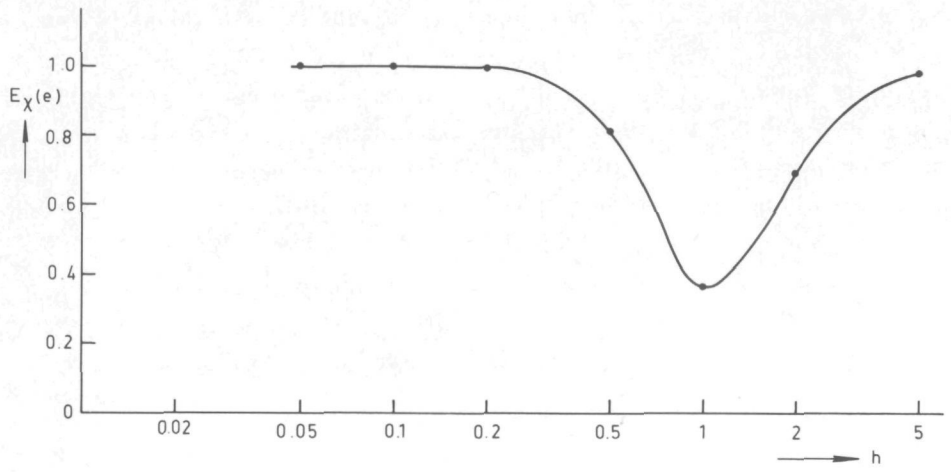


Fig. 3.16 The expected estimation error of one of the classes in a five dimensional classification problem as a function of the smoothing parameter of the Parzen estimation. Uniform kernels are used. m = 50.

h	$E_\chi(e)$
0.05	1.00 (0.00)
0.10	1.00 (0.00)
0.20	1.00 (0.00)
0.50	0.82 (0.02)
1.00	0.37 (0.03)
2.00	0.70 (0.03)
5.00	0.99 (0.01)

Table 3.8 The expected estimation error $E_\chi(e)$ of a five dimensional normal distribution using a Parzen estimation with uniform kernels and smoothing parameter h (m = 50). The presented values are the mean results of a Monte Carlo simulation. Between the brackets the computed standard deviations of the means are given.

of uniform kernels the density estimates are the same for the two classes if h is large enough, because then all learning objects contribute in the same way. The classification error approaches, therefore, by increasing h the a priori error min{c,1-c}, which is 0.5 in the presented example. For small values of h a similar effect exists. The two density estimates become zero almost everywhere, which results in the classification error approaching the a priori error. In that case normal kernels remain giving nonzero estimates everywhere.

From this experiment it can be concluded that in spite of the nearly identical estimation errors, the choice of the value of the smoothing parameter may be more critical for uniform kernels than for normal ones.

3.5 CONCLUDING REMARKS

The classification error made by an estimated discriminant function depends upon the accuracy of the density function estimates. From the examples in 3.4 it appeared, however, that this dependency differs from problem to problem, and can be very non-linear. An upper bound has been presented for the classification error which is expressed into the Bayes error and the estimation errors of the density functions. This bound expresses the worst thing that may happen: the complete estimation error works through into the classification error. In general it is highly improbable that this will occur. The advantage of the upper bound is that it can be computed, as we have shown, if only the family to which the class densities belong is known. If more detailed knowledge is available, such as in the case of normal distributions with equal covariance matrices, much lower figures for the expected classification error can be obtained.

The figures given in this chapter show how fast the classification error converges to ε by increasing sample size. From these figures it appears that the relation between ε^*, $\bar{\varepsilon}$, m and k can be roughly written as

$$\bar{\varepsilon} \approx \varepsilon^* + F\ (m/k) \tag{3.42}$$

(For the case of the general measurement space k has to be replaced by n). This is illustrated in fig. 3.17 where the data of the figures 3.4a, 3.4b and 3.6 are given as a function of m/k and m/n using a logarithmic scale for $\bar{\varepsilon} - \varepsilon^*$.

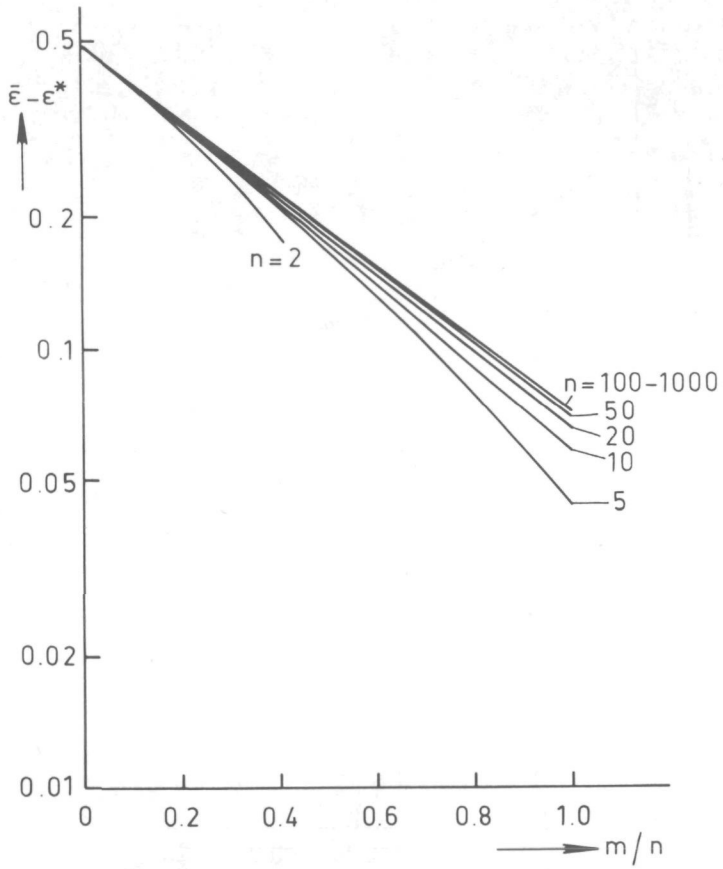


Fig. 3.17.a $\bar{\varepsilon} - \varepsilon^*$ *versus m/n for the data of fig. 3.4.a for several values of n.*

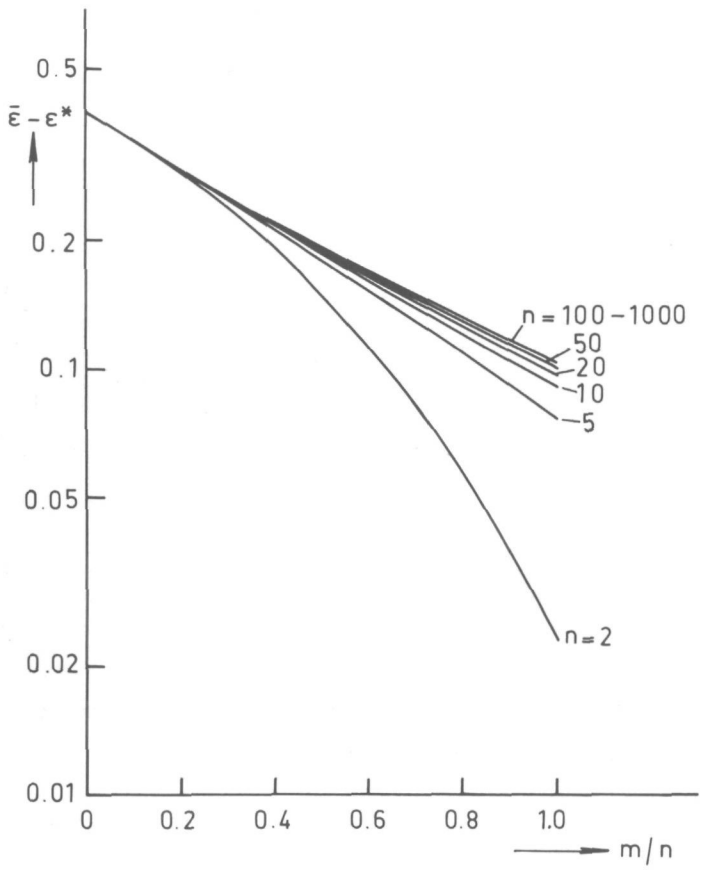


Fig. 3.17.b $\bar{\varepsilon} - \varepsilon^*$ *versus m/n for the data of fig. 3.4.b for several values of n.*

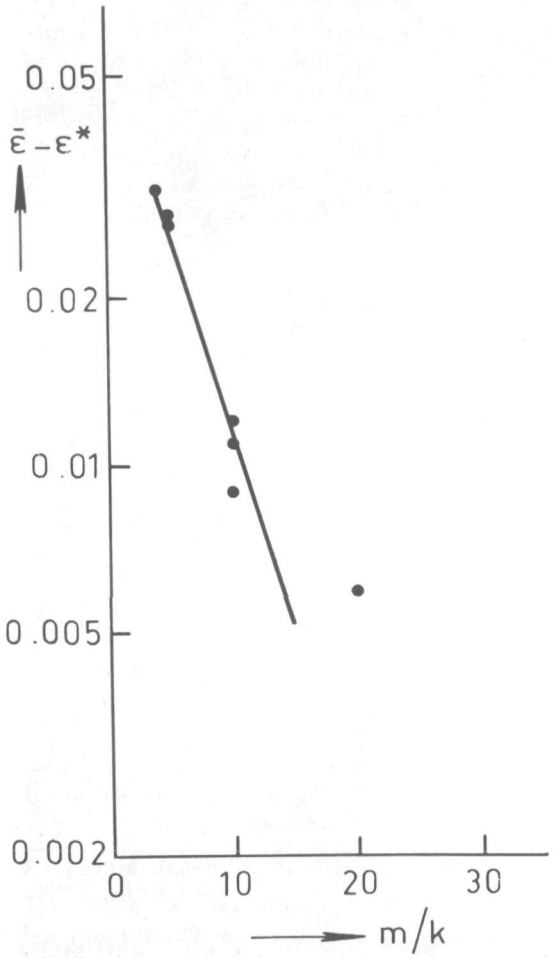


Fig. 3.17.c $\bar{\varepsilon} - \varepsilon^*$ *versus m/k for m/k > 4 for the data of table 3.3 (fig. 3.6). Points with* $\bar{\varepsilon} - \varepsilon^* < 0.003$ *are not given because of the relatively large variances.*

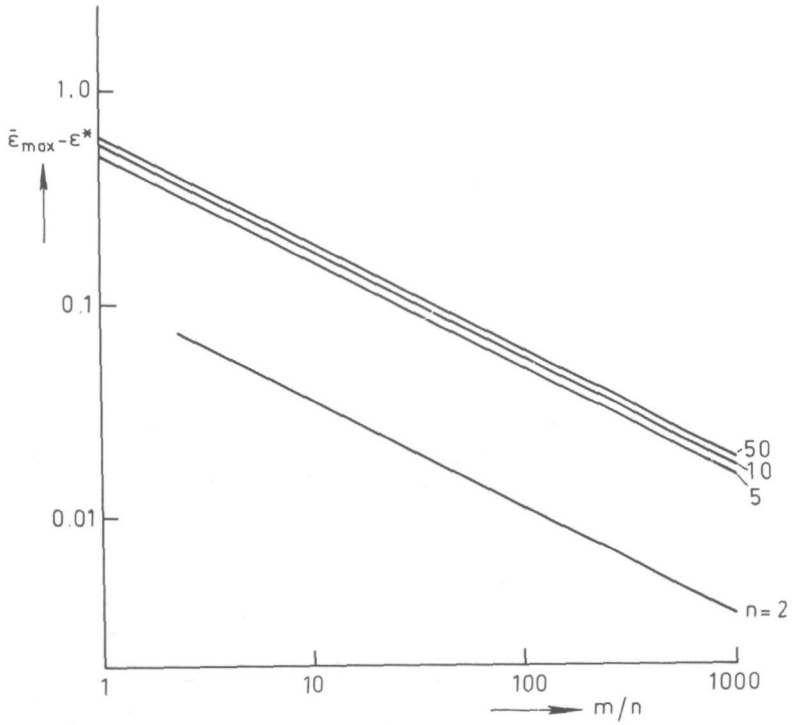


Fig. 3.18.a $\bar{\varepsilon}_{max} - \varepsilon^*$ *versus m/n for the data of fig. 3.3.a for several values of n.*

As these functions may be approximated by straight lines, this indicates that $F(\cdot)$ is for those cases an exponential function. The data of the upper bound in the general measurement space as given in fig. 3.3 shows a linear relation between $\bar{\varepsilon}_{max} - \varepsilon^*$ and m/n if for both a logarithmic scale is used, see fig. 3.18. This implies that the relation between the two quantities $\bar{\varepsilon}_{max} - \varepsilon^*$ and m/n is such that the one is the other raised to some power, except for some constants. This was already shown for larger values of n by (3.37). This illustrates that the convergence of the upper bound for $\bar{\varepsilon}$ is much slower than the exponential convergence of $\bar{\varepsilon}$ itself.

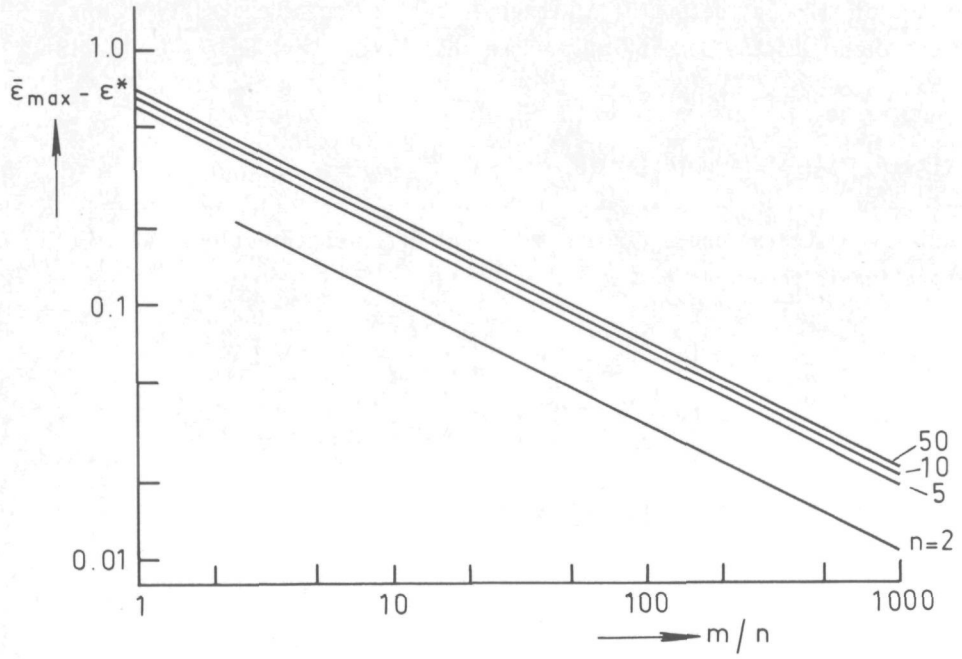


Fig. 3.18.b $\bar{\varepsilon}_{max} - \varepsilon^*$ *versus m/n for the data of fig. 3.3.b for several values of n.*

So far a restriction has been made to the two class case. For the multiclass case an upper bound similar to (3.18) can be found (for simplicity equal class probabilities will be assumed),

$$\varepsilon \leq \varepsilon^* + 2 \sum_{\ell=1}^{n_c} e_\ell / n_c \tag{3.43}$$

in which n_c is the number of classes. This follows immediately from (3.18) with the interpretation that class ℓ is class A, and all other classes together constitute class B. The maximum contribution of the estimation error of that class, e_ℓ, to the classification error ε is $2\, e_\ell / n_c$. Thus (3.43) gives the maximum classification error, given all estimation errors. For $\bar{\varepsilon}$ is found

$$\bar{\varepsilon} \leq \varepsilon^* + 2 \sum_{\ell=1}^{n_c} E_X(e_\ell) / n_c \tag{3.44}$$

A tighter bound can be derived as follows. In a multiclass problem several class densities may be estimated erroneously. For each object to be classified

only two of them may determine a wrong classification: the density of the correct class and the density of the one that takes over. So for each point $\underline{x}$ can be stated

$$\varepsilon(\underline{x}) \leq \varepsilon^*(\underline{x}) + 2 \max_{\ell} (e_\ell(\underline{x}))/n_c \tag{3.45}$$

in which $\varepsilon(\underline{x})$, $\varepsilon^*(\underline{x})$ and $e_\ell(\underline{x})$ are the local error contributions. After integration is found for ε

$$\varepsilon \leq \varepsilon^* + 2 \int_{\underline{x}} \max_{\ell} (e_\ell(\underline{x}))/n_c \, d\underline{x} \tag{3.46}$$

and for $\overline{\varepsilon}$

$$\overline{\varepsilon} \leq \varepsilon^* + 2 \, E_X\{\int_{\underline{x}} \max_{\ell} (e_\ell(\underline{x}))/n_c \, d\underline{x}\} \tag{3.47}$$

Especially when the number of classes is much more than two, the bound (3.47) may give a significant reduction compared with (3.44). The computation of (3.47), however, may be very difficult because a complicated function has to be integrated over $\underline{x}$.

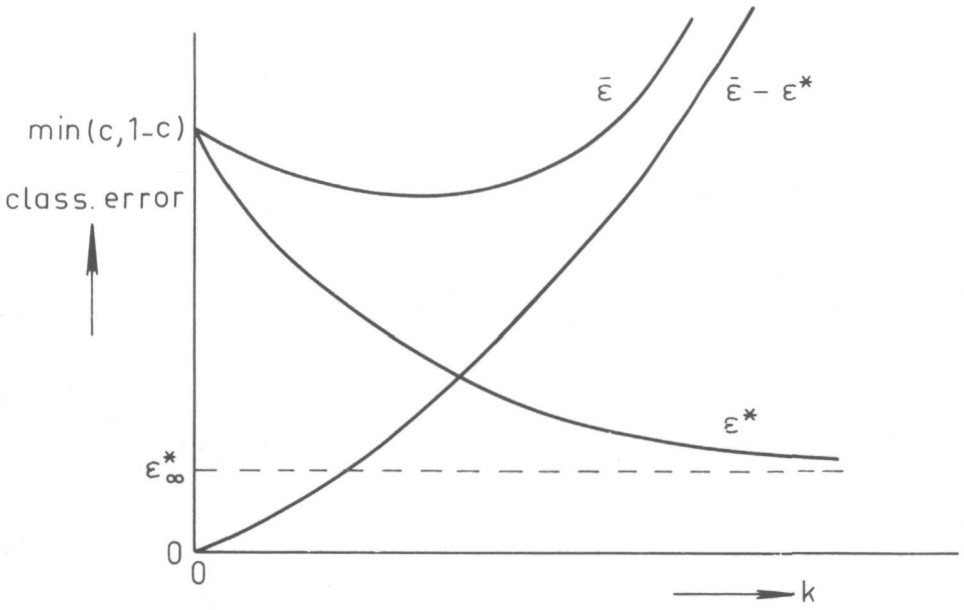


Fig. 4.1 $\bar{\varepsilon}$ *and* ε^* *as a function of the feature size.*

Chapter 4
FEATURE SIZE

4.1 THE PEAKING PHENOMENON

By considering fig. 3.6 it appears to be possible that for constant ε^* and constant sample size m the expected classification error $\bar{\varepsilon}$ is an increasing function of the feature size k. This seems to be counterintuitive, more features and worse results. The cause is that in spite of the fact that more features are available, no extra discriminating power is added, because the Bayes error has a constant value of ε^*. If this happens the new features are useless. They even cause worse results, because more parameters have to be estimated using the same number of objects.

Assume that the new features add some discriminating power, so ε^* is not constant but decreases with k and approaches asymptotically some value ε^*_∞. The difference between the expected classification error and the Bayes error $\bar{\varepsilon} - \varepsilon^*$, is, as explained in the previous chapter, often an increasing function of the expected estimation errors and thereby an increasing function of k. An exception has to be made for a very good new feature that makes $\bar{\varepsilon}$ as well as ε^* equal to zero. However, if the features are ranked in an order of decreasing discrimination power it often may happen that $\bar{\varepsilon} - \varepsilon^*$ increases with k because of the increasing expected estimation errors. The expected classification error $\bar{\varepsilon}$ may for that reason be a peaked function as indicated in fig. 4.1. This peaking phenomenon may also be observed in the classification error itself and in the mean classification error $\tilde{\varepsilon}$.

This chapter will be devoted to the peaking phenomenon. Some comments will be given on the literature and conditions will be presented under which peaking does not occur. Parts of this chapter have already been published in [15] and [16]. Examples of peaking are given by Ullmann [43] (classification of hand-printed numerals), Van Vark [46] (classification of human skeletal remains), Allais [3] (prediction), Van Ness and Simpson [44] (simulations of

discriminant analysis of normally distributed classes) and by Bouillon, Odell and Duran [4] (simulation of linear discriminant analysis of normally distributed classes). General discussions on feature size are given by Cover [9], Foley [22], Kanal and Chandrasekaran [25] and Raudys [35],[36].

The possibilities of peaking of ε, $\bar{\varepsilon}$ and $\tilde{\varepsilon}$ will now be treated shortly. Meanwhile a short introduction to the literature as well as to the other paragraphs of this chapter will be given.

Peaking of the classification error ε may always happen if the classes are not completely separable, i.e. if the class densities show some overlap. In that case it is for instance possible that the learning objects are so poor for a particular feature that a completely wrong picture of the relative positions of the classes is obtained as in fig. 3.2. The classification error ε may increase by such a feature. Another possibilities is that the class densities are completely identical for some feature while this cannot be detected from the learning objects. Such a feature will probably cause peaking of ε. A test on equality of the feature densities for the two classes will detect some bad features, but will not be able to avoid peaking completely.

From this point we will adopt the interpretation of Chandrasekaran and Jain [8] of the word peaking, which is more global than the one used just above. They call an increase of ε as a function of k only peaking if this increase is permanent. A local peak, caused by a single bad feature is therefore not a peak in this sense. Peaking of the classification error, in this interpretation, is still data dependent and cannot be proved from the data only. Even the use of a finite test set for making an estimate $\hat{\varepsilon}$ of ε will not do because peaking of $\hat{\varepsilon}$ does not necessarily coincide with peaking of ε.

The peaking phenomenon can in our set up be studied much better on the level of the expected classification error $\bar{\varepsilon}$, because in that case the density functions are assumed to be known. The results will have no direct practical value, because in a practical problem usually only learning sets are given. They show, however, what might happen and give some idea under what circumstances (sample size, feature size, density functions) peaking might be possible.

For studying the mean classification error $\tilde{\varepsilon}$ one is not restricted to a particular classification problem as one studies now the expectation of $\bar{\varepsilon}$ over a class of problems. This can be important when the density functions are given except for some parameters $\underline{\theta}$. The first published study on peaking of $\tilde{\varepsilon}$ was by Hughes [24], who used the model described in 3.2. In 4.2 Hughes' results

will be discussed as well as the comments given by Abend, Harley and Chandrasekaran [1].

After Hughes paper a number of papers appeared by Chandrasekaran and varying co-authors [5], [6], [7], [8], [25], which discussed several aspects of peaking. In particular the case of independent features drew their attention. For that case the peaking phenomenon is still more striking: new, independent information cannot be used because it causes worse results. In addition, the mathematical computations become in the case of independent features somewhat more feasible. For the case of independent binary features Chandrasekaran [5] showed that for a uniform parameter distribution peaking of $\tilde{\varepsilon}$ does not occur. In 4.3 we will give for the general case (not restricted to binary features) conditions for the parameter distribution and the estimators used, under which the mean classification error does not peak.

Chandrasekaran and Jain [7] constructed conditions under which the expected classification error $\bar{\varepsilon}$ does not peak. These conditions, however, appeared to be neither necessary nor sufficient as has been shown by a clarifying paper by Van Ness [45]. See also Duin [15] and Chandrasekaran and Jain [8]. These conditions, intended for preventing peaking of the expected classification error $\bar{\varepsilon}$, gave inspiration to some of the results for the mean classification error $\tilde{\varepsilon}$ presented in 4.3.

In 4.4 it will be shown and illustrated that one of the causes of peaking may be the choice of the estimators. If the right estimators are chosen peaking may sometimes be prevented. The knowledge, however, that has to enable us to make the right choice is often not available; suggestions for better choices are given for some cases.

The peaking phenomenon is not restricted to pattern recognition. Allais [3] showed that it exists in the related field of prediction. It also exists in regression analysis where it is easier to understand and computations are more simple. Because this field is rather out of the scope of this thesis we have restricted ourselves to a short discussion in appendix B.

4.2 DISCUSSION OF HUGHES' RESULTS

We will summarize here the results of Hughes [24] and of some comments on his paper made by Abend, Harley and Chandrasekaran [1] and present our own comments. Hughes' paper was the first to study the mean accuracy of statistical pattern recognizers. It resulted in curves for the mean error as a function of

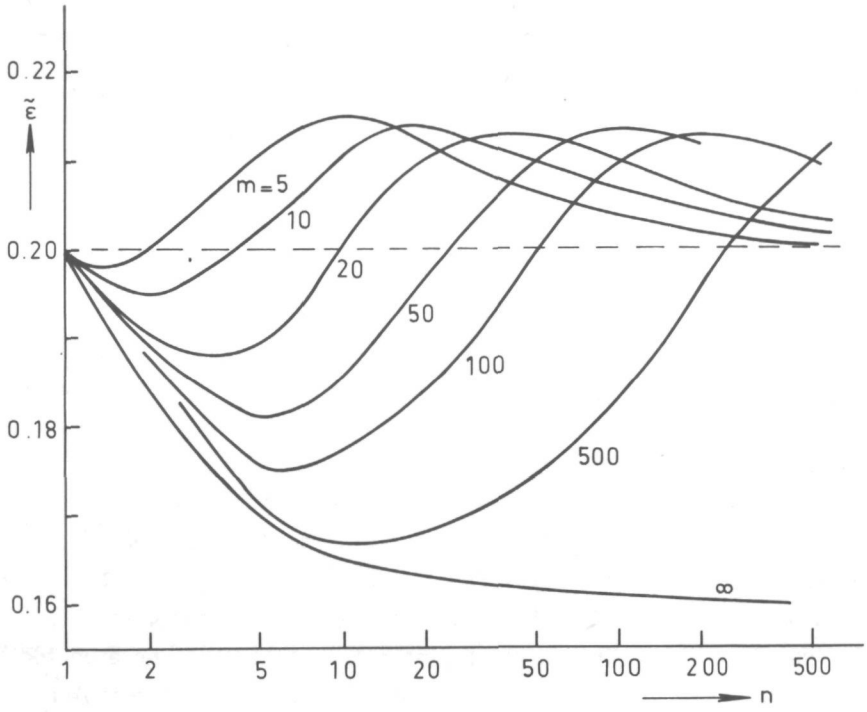


Fig. 4.2 The mean classification error $\tilde{\varepsilon}$, as computed by Hughes [24] as a function of measurement complexity n and sample size m. c = 0.2.

m	n_{opt}	$\tilde{\varepsilon}$
2	2	0,417
5	3	0,365
10	3	0,345
20	4	0,321
50	6	0,297
100	8	0,284
200	11	0,274
500	17	0,265
1000	23	0,261
∞	∞	0,250

Table 4.1 Optimum measurement complexity as a function of sample size and the corresponding mean classification error as computed by Hughes [24]. c = 0.5.

measurement complexity and sample size. Thereby was proved that peaking of the mean classification error existed.

The model of the measurement space introduced by Hughes is described in 3.2. The possible number of outcomes of a measurement x (also named cells) was called the *measurement complexity* and denoted by n. The probability that x is in cell j is indicated by

$$p_\ell^j = \text{Prob}\,(x \text{ in cell } j \mid x \in \text{class } \ell) \quad (j = 1,n;\ \ell = A,B) \tag{4.1}$$

The learning set consists of m objects for each class with cell frequencies q_A^j respectively q_B^j for cell j (j = 1,n), while $\sum_{j=1}^{n} q_A^j = \sum_{j=1}^{n} q_B^j = m$. Hughes made use of maximum likelihood estimators for p_A and p_B.

$$\hat{p}_\ell^j = \frac{q_\ell^j}{m} \quad (\ell = A,B) \tag{4.2}$$

Independent uniform distributions were used for the parameters which implies that $h(p_A^1, p_A^2, \ldots, p_A^n, p_B^1, p_B^2, \ldots, p_B^n)$ is equal for all parameter values under the restrictions

$$\sum_{j=1}^{n} p_A^j = 1$$

and (4.3)

$$\sum_{j=1}^{n} p_B^j = 1$$

The mean classification error $\tilde{\varepsilon}$, as given by (1.11), could be calculated analytically as a function of m, n and c. In fig. 4.2 Hughes' result is shown for c = 0.2. Two points are important to note. First the peaking phenomenon. After a certain value of n, called the *optimal measurement complexity*, $\tilde{\varepsilon}$ starts to increase. This value is a function of the sample size m, see table 4.1. Note that n = 32 is equivalent to a five dimensional binary feature space. Even for large sample sizes such as m = 1000, the optimal measurement complexity is very low.

The second point of interest that can be observed in fig. 4.2 is the fact that after peaking, the mean error continues to increase to values higher than the a priori probability of error of min{c, 1-c}. For these measurement complexities it would be better to classify all samples into the class with the highest a priori probability.

The comments of Abend, Harley and Chandrasekaran [1] concentrated on this second point. They state that this behaviour is caused by the fact that Hughes made use of maximum likelihood estimators instead of Bayes estimators, in spite of the fact that the parameter distributions were known. They prove that the following Bayes estimators correspond with the uniform distributions:

$$\hat{p}_{\ell}^{j} = \frac{q_{\ell}^{j}+1}{m+2} \qquad (j = 1,n;\ \ell = A,B) \tag{4.4}$$

They show that if use is made of these estimators ε behaves as is shown in fig. 4.3. This comment is correct, however, the fact still exists that if it is not known that the parameter distribution is uniform, one still might be confrontated with results as in fig. 4.2. The important thing to note for us is that in spite of the fact that the optimal estimators (4.4) are used peaking exists.

The reason why peaking exists in Hughes model is rather obvious, but it is not explicitly stated by Hughes or in the comments. The higher the model complexity, the more cells, and, because of the constant sample size, the more empty cells. An empty cell, without any learning object from class A or B, has to be allocated to the class with the highest a priori probability and gives a contribution to the error of cp_A^j if $c < 0.5$ and $(1-c)p_B^j$ if $c \geqslant 0.5$. This has to be summed over all empty cells. If the number of cells increases, this sum increases and approaches min{c, 1-c} because the fraction of non empty cells is reduced to zero.

The surprising feature of Hughes' results is not the existence of peaking, but rather the very low measurement complexity for which peaking occurs. This cannot be understood on the basis of empty cell considerations alone. Also the surprisingly large error contributions of the non empty cells have to be taken into account.

We will illustrate the importance of empty cells with an example of peaking in the expected error based on Hughes' model, which illustrates the influence of empty cells. Suppose the cell probabilities are given by

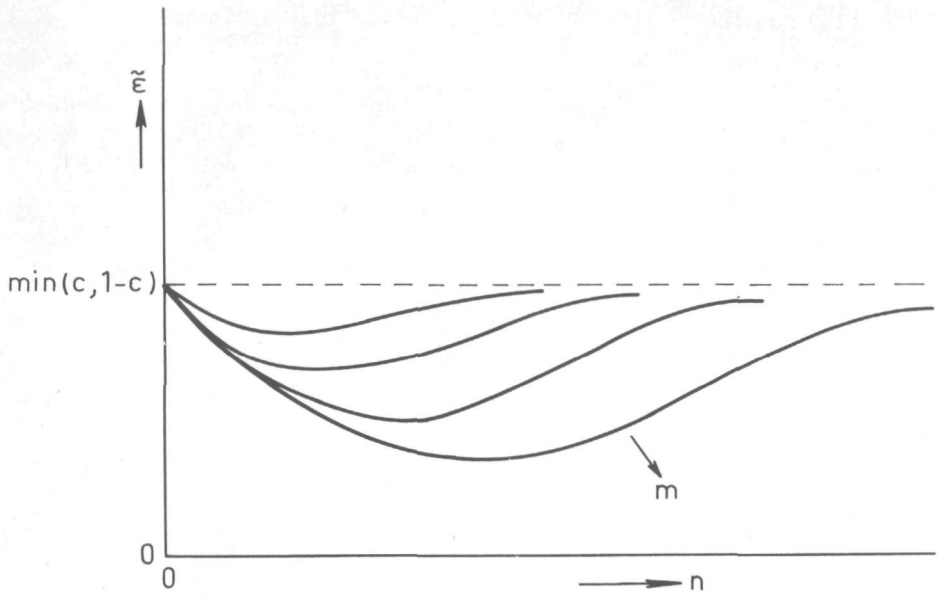


Fig. 4.3 The mean classification error $\tilde{\varepsilon}$ as a function of measurement complexity n and sample size m using Bayes estimators based on a uniform parameter distribution in Hughes' model.

$$p_A^j = 2/n, \quad p_B^j = 0 \quad (j = 1, n/2)$$

and (4.5)

$$p_A^j = 0, \quad p_B^j = 2/n \quad (j = n/2+1, n)$$

with n even. For the estimation of these probabilities the Bayes estimator (4.4) will be used. The expected error $\bar{\varepsilon}$ can be written, as follows from (1.8) and (1.10), supposing that $c = \frac{1}{2}$

$$\bar{\varepsilon} = \tfrac{1}{2}E_\chi\{\text{Prob}(\hat{S}(x) < 0 \mid x \in A, \chi)\} + \tfrac{1}{2}E_\chi\{\text{Prob}(\hat{S}(x) \geqslant 0 \mid x \in B, \chi)\} \tag{4.6}$$

The discriminant function for x in cell j is, using (4.4), given by

$$\hat{S}(x) = \tfrac{1}{2}\hat{p}_A^j - \tfrac{1}{2}\hat{p}_B^j = \tfrac{1}{2}\,\frac{q_A^j - q_B^j}{m+2} \quad (j = 1,n) \tag{4.7}$$

From (4.7) and (4.5) can be understood that the probability of $\hat{S}(x) < 0$ for $x \in A$ and the probability of $\hat{S}(x) > 0$ for $x \in B$ are zero. Expression (4.6) simplifies therefore to

$$\bar{\varepsilon} = \tfrac{1}{2}E_\chi\{\text{Prob}(\hat{S}(x) = 0 \mid x \in B, \chi)\} \tag{4.8}$$

$\hat{S}(x)$ can only be zero for points x with $q_B^j = 0$, which has a probability of $(1-p_B^j)^m$. So

$$\bar{\varepsilon} = \tfrac{1}{2}\sum_{j=1}^{n}\{(1-p_B^j)^m\, p_B^j\} \tag{4.9}$$

Substitution of (4.5) yields

$$\bar{\varepsilon} = \tfrac{1}{2}(1-2/n)^m \tag{4.10}$$

When the measurement complexity is raised to infinity one finds

$$\lim_{n\to\infty} \bar{\varepsilon} = \lim_{n\to\infty} \tfrac{1}{2}(1-2/n)^m = \tfrac{1}{2} \tag{4.11}$$

while for n=2 the value of $\bar{\varepsilon}$ is zero. Note that the classes do not overlap at all ($\varepsilon^* = 0, \forall\, n$). In this example the peaking is caused completely by the

empty cells. The expected error is the probability of finding an empty cell times the probability of an incorrect guess.

4.3 CONDITIONS FOR NO PEAKING OF $\tilde{\varepsilon}$

In this paragraph conditions are given under which, for any sample size, the mean classification error $\tilde{\varepsilon}$ over the class of independent feature distributions approaches zero monotonically if the feature size is raised to infinity.

The results of Hughes, described in 4.2 have been further investigated by Chandrasekaran et al. [5], [6], [7]. A different model was used in which $\tilde{\varepsilon}$ was studied as a function of the feature size instead of the measurement complexity. For the case of independent binary features Chandrasekaran [8] showed that $\tilde{\varepsilon}$ has no peaking and approaches zero if the feature size is raised to infinity. Like Hughes, he used a uniform distribution for the parameters.

We will present here more general conditions under which, for the case of independent features, the mean classification error will show no peaking. Under these conditions $\tilde{\varepsilon}$ will approach to zero if the feature size is raised to infinity. The distribution assumed for the parameters is not necessarily uniform. The proof is partly inspired by a paper by Chandrasekaran and Jain [7], which studies the peaking of the expected classification error $\bar{\varepsilon}$.

The estimate of the discriminant function $R(\underline{x})$ given by (1.3) will be written in a different way. Define

$$r^j = \log\{\hat{f}_A^j(x_j)\} - \log\{\hat{f}_B^j(x_j)\} \quad (j = 1, k) \tag{4.12}$$

and

$$d = \log\{(1-c)/c\} \tag{4.13}$$

$\hat{f}_\ell^j$ is the density estimate of feature j for class ℓ. $\hat{R}(\underline{x})$ is now given by

$$\hat{R}(\underline{x}) = \sum_{j=1}^{k} r^j - d \tag{4.14}$$

It is assumed that the features are independently distributed. For further calculations it will be necessary for the true feature densities $f_A^j(\cdot)$ and $f_B^j(\cdot)$ to be such that

$$f^j_\ell(x_j|\underline{\theta}^j_\ell) = f(x_j|\underline{\theta}^j_\ell) \qquad (j = 1,k;\ \ell = A,B) \tag{4.15}$$

This implies that all feature distributions are of the same type, e.g. normal distributions. Let the joint parameter density of $\underline{\theta}_A$ and $\underline{\theta}_B$ be given by $h(\underline{\theta}_A,\underline{\theta}_B)$, from which the parameters are drawn independently, be independent of j. The variables r^j are thus random variables with respect to $\underline{\theta}^j_A$, $\underline{\theta}^j_B$, $\underline{x}$ and χ. Since the parameter sets are selected independently and the classes have independent features the variables r^j can be interpreted as independent identically distributed random variables. The mean classification error (1.11) is now given by, as follows from (1.8)

$$\tilde{\varepsilon} = c\ \text{Prob}\Big(\sum_{j=1}^{k} r^j < d\Big) + (1-c)\ \text{Prob}\Big(\sum_{j=1}^{k} r^j \geqslant d\Big) \tag{4.16}$$

If the variance of a variable r^j is bounded and if the expectation of r^j is positive, then the probability that a sum of k of those variables is larger than a fixed constant d approaches one by increasing k. From now on it is assumed that the variance of r^j is bounded. Sufficient conditions under which the probabilities in (4.16) go to zero are therefore

$$E_{\underline{\theta}}\ E_{\chi}\ E_{\underline{x}\in A}(r^j) > 0 \tag{4.17}$$

and

$$E_{\underline{\theta}}\ E_{\chi}\ E_{\underline{x}\in B}(r^j) < 0 \tag{4.18}$$

which has to be true for each j. Since it is assumed that the density of the parameters, $h(\underline{\theta}_A, \underline{\theta}_B)$, is independent of j, the r^j's have for all j the same distribution, because of (4.15). $\tilde{\varepsilon}$ will therefore approach to zero monotonically.

It will now be proved that (4.17) and (4.18) are fulfilled if $h(\underline{\theta}_A, \underline{\theta}_B)$ satisfies

$$h(\underline{\theta}_A, \underline{\theta}_B) = h(\underline{\theta}_B, \underline{\theta}_A) \tag{4.19}$$

for each $\underline{\theta}_A$ and $\underline{\theta}_B$ and

$$\int_{\underline{\theta}_A \neq \underline{\theta}_B} h(\underline{\theta}_A, \underline{\theta}_B)\, d\underline{\theta}_A\, d\underline{\theta}_B > 0 \tag{4.20}$$

and if Q, defined by

$$Q(x_j, \underline{\theta}_A, \underline{\theta}_B) = E_X(r^j) \tag{4.21}$$

satisfies

$$Q(x_j, \underline{\theta}_A, \underline{\theta}_B) > 0 \quad \text{if} \quad f(x_j|\underline{\theta}_A) > f(x_j|\underline{\theta}_B) \tag{4.22}$$

$$Q(x_j, \underline{\theta}_A, \underline{\theta}_B) = 0 \quad \text{if} \quad f(x_j|\underline{\theta}_A) = f(x_j|\underline{\theta}_B) \tag{4.23}$$

$$Q(x_j, \underline{\theta}_A, \underline{\theta}_B) < 0 \quad \text{if} \quad f(x_j|\underline{\theta}_A) < f(x_j|\underline{\theta}_B) \tag{4.24}$$

Note that

$$Q(x_j, \underline{\theta}_A, \underline{\theta}_B) = -\, Q(x_j, \underline{\theta}_B, \underline{\theta}_A) \tag{4.25}$$

because of (4.21), (4.15) and (4.12). For the proof, condition (4.17) will be written as

$$\int_{x_j} \int_{\underline{\theta}_A} \int_{\underline{\theta}_B} Q(x_j, \underline{\theta}_A, \underline{\theta}_B)\, f(x_j|\underline{\theta}_A)\, h(\underline{\theta}_A, \underline{\theta}_B)\, d\underline{\theta}_A\, d\underline{\theta}_B\, dx_j > 0 \tag{4.26}$$

The integral over $\underline{\theta}_A$ and $\underline{\theta}_B$ can be split into a sum of three terms, one for the region with $f(x_j|\underline{\theta}_A) > f(x_j|\underline{\theta}_B)$, one for the region with $f(x_j|\underline{\theta}_A) < f(x_j|\underline{\theta}_B)$ and one for the region with $f(x_j|\underline{\theta}_A) = f(x_j|\underline{\theta}_B)$. The integral over this last region is zero because of (4.23). If $\underline{\theta}_A$ and $\underline{\theta}_B$ are interchanged in the integral over the region with $f(x_j|\underline{\theta}_A) < f(x_j|\underline{\theta}_B)$ one gets, using (4.19) and (4.25)

$$\int_{x_j} \iint_{f(x_j|\underline{\theta}_A) > f(x_j|\underline{\theta}_B)} Q(x_j, \underline{\theta}_A, \underline{\theta}_B)\{f(x_j|\underline{\theta}_A) - f(x_j|\underline{\theta}_B)\} h(\underline{\theta}_A, \underline{\theta}_B)\, d\underline{\theta}_A\, d\underline{\theta}_B\, dx_j > 0 \tag{4.27}$$

All factors in the integrand are positive because of (4.20) and (4.22), which

causes that this condition, and thereby (4.17), are satisfied. In the same way it can be proved that the presented conditions (4.19), (4.20) and (4.22) - (4.24) are sufficient for (4.18) to be true.

The conditions (4.17) and (4.18) guarantee, for the case of independent distributions, that the mean classification error approaches zero monotonically. These conditions differ slightly from the ones given by Chandrasekaran and Jain [7], who were interested in $\bar{\varepsilon}$ instead of $\tilde{\varepsilon}$. It appeared that the conditions of Chandrasekaran and Jain are inaccurate. This will be treated in the next paragraph in more detail.

We proved that our conditions are fulfilled if (4.19) - (4.24) are satisfied. The condition (4.19) is probably the most demanding one. It requires that the chances of finding a feature with parameters $\underline{\theta}_A$ for class A and $\underline{\theta}_B$ for class B are as great as finding a feature with parameters $\underline{\theta}_B$ for A and $\underline{\theta}_A$ for B.

Condition (4.20) simply demands that the classes differ in their statistical behaviour. This is trivial, because otherwise a worthwhile separation is impossible. The conditions (4.19) and (4.20) include the assumption of uniform distributions for $\underline{\theta}_A$ and $\underline{\theta}_B$ over the same interval.

The conditions (4.22) - (4.24) demand that the expected value (over all learning sets) of the estimated discriminant function has the same sign as the optimal one. In the next paragraph an example of this will be given.

4.4 INFLUENCE OF THE ESTIMATORS

In the previous paragraph a class of estimators is defined implicitly that fulfils the presented conditions for no peaking of $\tilde{\varepsilon}$ if the parameter density function satisfied (4.19) and (4.20). Independent class distributions were assumed. Here it will be shown that for known parameter density functions an estimator can be constructed which always prevents peaking of $\tilde{\varepsilon}$. This estimator is for the case of independent class distributions a member of the class mentioned above.

The case of independent features, starting from the conditions (4.17) and (4.18) will be considered first. Condition (4.17) is after substitution of (4.12) equivalent to

$$\int_{\underline{\theta}} \int_{\chi} \int_{x_j} \{\log\{\hat{f}_A^j(x_j)\} - \log\{\hat{f}_B^j(x_j)\}\} \, f_A^j(x_j|\underline{\theta}_A) g(\chi|\underline{\theta}) h(\underline{\theta}) dx_j d\chi d\underline{\theta} > 0 \qquad (4.28)$$

The estimates $\hat{f}_A^j(x_j)$ and $\hat{f}_B^j(x_j)$ are computed from χ alone. So if the order of integration is changed and integration over $\underline{\theta}$ is carried out first, (4.28) becomes

$$\int_{\chi} \int_{x_j} v_A^j(x_j,\chi)\{\log\{\hat{f}_A^j(x_j)\} - \log\{\hat{f}_B^j(x_j)\}\} \, dx_j d\chi > 0 \qquad (4.29)$$

in which

$$v_A^j(x_j,\chi) = \int_{\underline{\theta}} f_A^j(x_j|\underline{\theta}_A) \; g(\chi|\underline{\theta}) \; h(\underline{\theta}) \; d\underline{\theta} \qquad (4.30)$$

is the marginal density of $x_j \in A$ and χ. Similar expressions can be found for class B starting from (4.18). Choose as density estimator

$$\hat{f}_\ell^j(x_j) = \frac{v_\ell^j(x_j,\chi)}{g(\chi)} = v_\ell^j(x_j|\chi) \qquad (\ell = A,B) \qquad (4.31)$$

in which $g(\chi)$ is the marginal density of χ

$$g(\chi) = \int_{\underline{\theta}} g(\chi|\underline{\theta}) \; h(\underline{\theta}) \; d\underline{\theta}$$

For the calculation of the estimator (4.31) the following factors have to be known: the parameter distribution $h(\underline{\theta})$, the functional form of $f_\ell^j(\cdot)$ and a learning set χ. Knowledge of the parameters $\underline{\theta}_A$ and $\underline{\theta}_B$ themselves is of course not necessary.

After some calculations condition (4.29) can now be written as

$$\int_{\chi} \int_{x_j} v_A^j(x_j,\chi) \; \log\{v_A^j(x_j,\chi)\} \; dx_j \; d\chi$$

$$- \int_{\chi} \int_{x_j} v_A^j(x_j,\chi) \; \log\{v_B^j(x_j,\chi)\} \; dx_j \; d\chi > 0 \qquad (4.32)$$

which is always satisfied unless $v_A^j(x_j,\chi) = v_B^j(x_j,\chi)$, $\forall(x_j,\chi)$. This can be understood after realizing that $\int_x a(x) \log\{b(x)\}dx$ in which $a(x)$ and $b(x)$ are density functions , is maximum for $b(x) = a(x)$, $\forall x$ (e.g. see Kullback [26]). It is easy to verify that condition (4.18) is also satisfied by choosing (4.31) as estimator. This completes the proof that the estimator (4.31) results by satisfying (4.17) and (4.18), in a non-peaking mean error $\tilde{\varepsilon}$.

These estimates are, due to (4.30), based on the known parameter density $h(\underline{\theta})$. If this density is not exactly known, as in many practical situations, it is not possible to make $\hat{f}_\ell^j(\cdot)$ exactly equal to $v_\ell^j(\cdot)$, but (4.29) may still hold. If $\hat{f}_\ell^j(\cdot)$ differs too much from $v_\ell^j(\cdot)$ no guarantee for avoiding peaking exists.

Note that the estimator (4.31) for $\ell = A$ as well as $\ell = B$ depends upon the entire learning set χ. In practice these kinds of estimators are rather unusual.

An example, which is a special case of the above proof, is given by Chandrasekaran [5]. For independent binary features with a uniform parameter distribution, he proves that the mean error does not peak in the case of Bayes estimates based on a uniform parameter distribution. In appendix C it is proved that if maximum likelihood estimators were used the mean error would show peaking.

If the features are not independent the conditions (4.17) and (4.18) cannot be used and a somewhat different approach has to be made. From (1.11) and (1.9) it follows that for $\tilde{\varepsilon}$ can be written

$$\tilde{\varepsilon} = \int_{\underline{\theta}} \iint_{(\underline{x},\chi)\,|\,\hat{S}(\underline{x},\chi)<0} c\, f_A(\underline{x}|\underline{\theta}_A)\, g(\chi|\underline{\theta})\, h(\underline{\theta})\, d\underline{x}\, d\chi\, d\underline{\theta} +$$

$$\int_{\underline{\theta}} \iint_{(\underline{x},\chi)\,|\,\hat{S}(\underline{x},\chi)\geq 0} (1-c)\, f_B(\underline{x}|\underline{\theta}_B)\, g(\chi|\underline{\theta})\, h(\underline{\theta})\, d\underline{x}\, d\chi\, d\underline{\theta} \qquad (4.33)$$

$S(\underline{x},\chi)$ is the discriminant function based upon a learning set χ. This function has to be defined in such a way that (4.33) is minimum. After interchanging the order of intergration, one finds

$$\tilde{\varepsilon} = \iint_{(\underline{x},\chi)\,|\,\hat{S}(\underline{x},\chi)<0} c\ v_A(\underline{x},\chi)\ d\underline{x}\ d\chi + \iint_{(\underline{x},\chi)\,|\,\hat{S}(\underline{x},\chi)\geq 0} (1-c)\ v_B(\underline{x},\chi)\ d\underline{x}\ d\chi \tag{4.34}$$

in which $v_\ell(\underline{x},\chi)$ is given by

$$v_\ell(\underline{x},\chi) = \int_{\underline{\theta}} f_\ell(\underline{x}|\underline{\theta}_\ell)\ g(\chi|\underline{\theta})\ h(\underline{\theta})\ d\underline{\theta} \qquad (\ell = A,B) \tag{4.35}$$

which is the mean joint density of $\underline{x} \in$ class ℓ and χ. If the second term in (4.34) is expressed in an integration over the complementary region, the expression for $\tilde{\varepsilon}$ simplifies to

$$\tilde{\varepsilon} = \iint_{(\underline{x},\chi)\,|\,\hat{S}(\underline{x},\chi)<0} \{c\ v_A(\underline{x},\chi) - (1-c)\ v_B(\underline{x},\chi)\}\ d\underline{x}\ d\chi + 1-c \tag{4.36}$$

which is minimum if all points in which the integrand is negative are used for the integration region. So

$$\hat{S}(\underline{x},\chi) = c\ v_A(\underline{x},\chi) - (1-c)\ v_B(\underline{x},\chi) \tag{4.37}$$

This can be used as discriminant function if the parameter distribution $h(\underline{\theta})$ is known. This discriminant function is the optimal one in the sense that it minimizes $\tilde{\varepsilon}$. Peaking of $\tilde{\varepsilon}$ will now be prevented. This can be understood after realizing that the (k+1)-dimensional feature space R_{k+1} contains the k-dimensional feature space R_k as a subspace. Thus each discriminant function in R_k is also a discriminant function in R_{k+1} by giving a zero weight to the new feature. The optimal discriminant function in R_{k+1} is as good as or better than the optimal discriminant function in R_k because the latter is available in R_{k+1}. Using these optimal discriminant functions guarantees therefore the absence of peaking.

The above results are in apparent contradiction with the results of Hughes as described in 4.2. There too the optimal discriminant function based on the known parameter distribution was used. However, in that case peaking still existed. This was caused by the different model in which the discriminant function in the space with measurement complexity n had no meaning, and could therefore not be used, in the space with measurement complexity n+1. So the results of an optimal discriminant function do not hold in a space with higher complexity and peaking may happen.

The result of no peaking obtained by using (4.37) is valid for the case of dependent features. Note that applying it to independent features gives the same discrimination as obtained by using the estimator (4.31) because

$$\prod_{j=1}^{k} v_{\ell}^{j}(x_j,\chi) = v_{\ell}(\underline{x},\chi) \qquad (\ell = A,B) \tag{4.38}$$

as follows from (4.30) and (4.35).

The optimal discriminant function for known $h(\underline{\theta})$ (4.37) is after substitution of (4.35) identical to the discriminant function (2.13) introduced in chapter 2. The above therefore proves that the Bayes estimator (2.14) is optimal, in the sense that it minimizes the mean classification error.

In chapter 2 an example of binary features was presented which will be considered here in more detail. A multivariable independent binary distribution is assumed. The density for a single feature is given by

$$f_{\ell}^{j}(x_j) = (p_{\ell}^{j})^{x_j} (1 - p_{\ell}^{j})^{1-x_j} \qquad (\ell = A,B;\ j = 1,k)$$

The three kinds of estimators (2.18), (2.20) and (2.22) are compared by computing $\tilde{\varepsilon}$ for a uniform parameter density $h(p_A,p_B)$ on the line $p_A = 1-p_B$ and a zero density elsewhere.

1) Maximum likelihood estimator for the density estimate of feature j

$$\hat{f}_{\ell}^{j}(x_j) = \left(\frac{n_{\ell}^{j}}{m}\right)^{x_j}\left(1 - \frac{n_{\ell}^{j}}{m}\right)^{1-x_j} \qquad (\ell = A,B) \tag{4.39}$$

Substitution into (4.12) yields

$$r^{j} = x_j\{\log n_A^j/n_B^j\} + (1-x_j)\log\{(m-n_A^j)/(m-n_B^j)\} \tag{4.40}$$

As n_A and n_B may have the values 0 and m, the expectation over the learning set is not defined. The conditions (4.17) and (4.18) cannot be applied for that reason. In appendix C is shown, however, that peaking exists by using (4.39) for density estimation.

2) The Bayes estimator constructed by assuming a uniform parameter density for each p_{ℓ}^{j}, which is the marginal density of $h(p_A,p_B)$.

$$\hat{f}_{\ell}^{j}(x_j) = \left(\frac{n_{\ell}^{j}+1}{m+2}\right)^{x_j} \left(1 - \frac{n_{\ell}^{j}+1}{m+2}\right)^{1-x_j} \tag{4.41}$$

Substitution into (4.12) yields

$$r^j = x_j \log\{(n_A^j+1)/(n_B^j+1)\} + (1-x_j) \log\{(m-n_A^j+1)/(m-n_B^j+1)\} \tag{4.42}$$

In appendix C it is proved that this r^j fulfils the conditions (4.17), (4.18) and (4.22) - (4.24). It can be easily verified that the given parameter distribution satisfies (4.19) and (4.20). Hereby it is clear that the estimator (4.41) shows no peaking in the presented example.

3) The Bayes estimator using the true parameter distribution, which is, as is shown in chapter 2

$$\hat{f}_{A}^{j}(x_j) = \left(\frac{m+n_A^j-n_B^j+1}{2m+2}\right)^{x_j} \left(1 - \frac{m+n_A^j-n_B^j+1}{2m+2}\right)^{1-x_j} \tag{4.43}$$

for class A and

$$\hat{f}_{B}^{j}(x_j) = \left(\frac{m+n_B^j-n_A^j+1}{2m+2}\right)^{x_j} \left(1 - \frac{m+n_B^j-n_A^j+1}{2m+2}\right)^{1-x_j} \tag{4.44}$$

for class B. These are, as indicated earlier in this paragraph the optimal estimators, because they are based on the true parameter density. The mean classification error will, therefore, show no peaking by using these estimators.

Some values of $\tilde{\varepsilon}$, for the three presented estimators, are given in table 4.2. These values are based on an exact computation. For large sample sizes and feature sizes a Monte Carlo procedure would have to be used, which is not accurate enough to show the difference between the results of the second and the third estimator. The results show that peaking exists when the maximum likelihood estimator is used. The difference between the two Bayes estimators appears to be rather small for the example presented. This indicates that the non-optimal estimator (4.41) is a robust one.

We will now consider the role of the estimator in the peaking of the expected classification error $\bar{\varepsilon}$. This has, for independent features, recently been treated by Chandrasekaran and Jain [7], in a comment on their paper made by Van Ness [45] and in the authors reply [8]. The results are

k	m	$\tilde{\varepsilon}$ in %		
		estimators used		
		1	2	3
2	0	50.0	50.0	50.0
2	1	33.3	27.8	27.8
2	2	26.0	23.3	23.3
2	3	23.0	21.5	21.4
2	4	21.4	20.5	20.4
2	5	20.4	19.9	19.7
2	6	19.8	19.4	19.2
3	0	50.0	50.0	50.0
3	1	37.0*	23.1	23.1
3	2	26.5*	18.7	18.6
3	3	21.6	17.0	16.7

Table 4.2 Values of $\tilde{\varepsilon}$ (in %) for the presented example (see text) for feature size k and sample size m. The three estimators used are

1. *The maximum likelihood estimator (4.39)*
2. *The Bayes estimator (4.41)*
3. *The Bayes estimators (4.43) and (4.44)*

* *These values show an increase after the addition of a new feature (peaking).*

summarized below, together with our own comments.

Chandrasekaran and Jain [7] presented the following conditions for no peaking of $\bar{\varepsilon}$

$$\lim_{k\to\infty} E_{\chi}\, E_{\underline{x}\in A}\left\{\sum_{j=1}^{k} r^j\right\} = \infty \tag{4.45}$$

$$\lim_{k\to\infty} E_{\chi}\, E_{\underline{x}\in B}\left\{\sum_{j=1}^{k} r^j\right\} = -\infty \tag{4.46}$$

in which r^j is defined as in (4.12). They state that if these conditions are fulfilled $\bar{\varepsilon}$ shows no peaking. Their argument is that the distribution of the sum of the independent random variables r^j becomes normal because of the central limit theorem. If the mean of that distribution goes to infinity for increasing k and the variances of the r^j's are sufficiently well behaved, then the expectation of the sum devided by its standard deviation also goes to infinity and complete separation between the classes becomes possible, so $\bar{\varepsilon} \to 0$ for $k \to \infty$.

Van Ness [45] showed that the conditions (4.45) and (4.46) are neither necessary nor sufficient since the central limit theorem is not valid for the general situation described and the standard deviation may approach infinity as fast as the mean does.

In their reply Chandrasekaran and Jain [8] admit these imperfections and give new, sufficient conditions that do not depend on the applicability of the central limit theorem. These conditions are

$$\lim_{k\to\infty} \frac{\sum_{j=1}^{k} E_{\chi}\, E_{\underline{x}\in A}(r^j)}{\{\sum_{j=1}^{k} \mathrm{Var}_{\chi}\, \mathrm{Var}_{\underline{x}\in A}(r^j)\}^{\frac{1}{2}}} = \infty \tag{4.47}$$

and

$$\lim_{k\to\infty} \frac{\sum_{j=1}^{k} E_{\chi}\, E_{\underline{x}\in B}(r^j)}{\{\sum_{j=1}^{k} \mathrm{Var}_{\chi}\, \mathrm{Var}_{\underline{x}\in B}(r^j)\}^{\frac{1}{2}}} = -\infty \tag{4.48}$$

The notations $\mathrm{Var}_{\chi}\mathrm{Var}_{\underline{x}\in A}(r^j)$ and $\mathrm{Var}_{\chi}\mathrm{Var}_{\underline{x}\in B}(r^j)$ stand for the variance of r^j due to the random choices of χ and $\underline{x}$. The variables r^j, given by (4.12), are functions of the density estimates of the classes. The expectations and the variances in (4.47) and (4.48) are therefore estimator dependent. Thus the choice of the estimator is relevant for avoiding peaking. In particular estimators with large or unbounded variances have to be avoided. For instance estimators for which the density estimates in (4.12) can be arbitrarily small or even zero (e.g. maximum likelihood estimators for binary features) cause large or unbounded variances of r^j and may thereby cause peaking. In appendix C it is shown that for the case of independent binary features maximum likelihood estimators nearly always result in peaking. The case of Bayes estimators based upon a uniform parameter distribution, as given by (4.41) is investigated by Chandrasekaran and Jain [7] and [8]. They found that in certain small regions of the parameter space peaking occurs, while in the other regions peaking is avoided. This illustrates the importance of the choice of the estimator.

Finally we will make some remarks on James-Stein estimators. These estimators can be used when the same parameters have to be estimated for a number of distributions, each represented by their own learning set. This is exactly

the situation in the case of independent features. The main supporters of the James-Stein estimators, Efron and Morris [18], [19], [20], [21] state that in some cases the estimators approach the Bayes estimator based on the true parameter distribution if the number of distributions goes to infinity. This would imply that for independent features peaking could be avoided in those cases. This point has not yet been further investigated.

4.5 DISCUSSION ON THE PEAKING PHENOMENON

The cause of peaking of the expected classification error $\bar{\varepsilon}$ will be discussed first, starting with some results of the previous paragraph. The conditions (4.47) and (4.48) can be generalized, using (4.14) as

$$\lim_{k\to\infty} \frac{E_X\, E_{\underline{x}\in A}(\hat{R}(\underline{x}))}{\{\mathrm{Var}_X\, \mathrm{Var}_{\underline{x}\in A}(\hat{R}(\underline{x}))\}^{\frac{1}{2}}} = \infty \tag{4.49}$$

and

$$\lim_{k\to\infty} \frac{E_X\, E_{\underline{x}\in B}(\hat{R}(\underline{x}))}{\{\mathrm{Var}_X\, \mathrm{Var}_{\underline{x}\in B}(\hat{R}(\underline{x}))\}^{\frac{1}{2}}} = -\infty \tag{4.50}$$

Van Ness [45] showed, using the Chebyshev inequality, that these conditions are also in the case of dependent features sufficient for avoiding peaking of $\bar{\varepsilon}$, because they guarantee that

$$\lim_{k\to\infty} \mathrm{Prob}(\hat{R}(\underline{x}) < 0 | \underline{x}\in A) = 0 \tag{4.51}$$

and

$$\lim_{k\to\infty} \mathrm{Prob}(\hat{R}(\underline{x}) \geqslant 0 | \underline{x}\in B) = 0 \tag{4.52}$$

On the other hand, if $\hat{R}(\underline{x})$ has an asymptotically symmetric distribution (e.g. normal), sufficient conditions for the existence of peaking are

$$\lim_{k\to\infty} \frac{E_X\, E_{\underline{x}\in A}(\hat{R}(\underline{x}))}{\{\mathrm{Var}_X\, \mathrm{Var}_{\underline{x}\in A}(\hat{R}(\underline{x}))\}^{\frac{1}{2}}} = 0 \tag{4.53}$$

and

$$\lim_{k\to\infty} \frac{E_\chi \, E_{\underline{x}\in B}(\hat{R}(\underline{x}))}{\{Var_\chi \, Var_{\underline{x}\in B}(\hat{R}(\underline{x}))\}^{\frac{1}{2}}} = 0 \tag{4.54}$$

In that case these conditions guarantee

$$\lim_{k\to\infty} Prob(\hat{R}(\underline{x}) < 0 | \underline{x}\in A) = \tfrac{1}{2} \tag{4.55}$$

and

$$\lim_{k\to\infty} Prob(\hat{R}(\underline{x}) \geqslant 0 | \underline{x}\in B) = \tfrac{1}{2} \tag{4.56}$$

The four conditions (4.49), (4.50), (4.53) and (4.54) show the importance of the expectation of $\hat{R}(\underline{x})$ compared with its standard deviation. In the case of independent features which are identically distributed, both the expectation and the variance of $\hat{R}(\underline{x})$ increase with order k. By this (4.49) and (4.50) are fulfilled and peaking is avoided with the following exception. If an increasing number of identically distributed independent features are added which have no discriminating power ($E_\chi \, E_{\underline{x}\in A}(r^j) = E_\chi \, E_{\underline{x}\in B}(r^j) = 0$) then $\hat{R}(\underline{x})$ has an assymptotic normal distribution and (4.53) and (4.54) are fulfilled. Generally, if the features are not identically distributed and worse and worse features are added, the standard deviation of $\hat{R}(\underline{x})$ may grow faster than its expectation and peaking becomes possible.

A fast increasing standard deviation of $\hat{R}(\underline{x})$ can be caused by the estimators used, such as the maximum likelihood estimator in the case of binary features, or by the class densities of the classification problem involved. An example of the latter is given by van Ness [44]. He proves that the discrimination of two normally distributed classes with unknown expectations and known variances produces peaking if the variances of the newly added features increase fast enough with the feature size.

So we find three causes of peaking of $\bar{\varepsilon}$, which are highly interrelated.

1. The choice of bad estimators.
2. Fast increasing variance of $\hat{R}(\underline{x})$.
3. Slowly increasing expectation of $\hat{R}(\underline{x})$.

Now the classification error ε will be considered. This error peaks whenever $\bar{\varepsilon}$ does, except if the learning set is, by accident, so good that the effects mentioned above are avoided. In general this will only delay peaking somewhat, because each new feature introduces new estimation errors. In the same way a bad learning sets may cause an early peaking.

Peaking of the mean classification error $\tilde{\varepsilon}$ has the same three causes as peaking of $\bar{\varepsilon}$, because here a number of problems is averaged of which each individually may show peaking. It completely depends on the weights for the problems in the parameter distribution whether $\tilde{\varepsilon}$ shows peaking or not.

As has been stated, the choice of the estimator is of great importance for avoiding peaking. The question, what is a good estimator, is, however, hard to answer in general. The answer is problem dependent and can only be given if some knowledge on the parameter values is available. So here the same problem as everywhere else in statistics is met: good estimators cannot be constructed without a priori knowledge on the parameter values.

In this chapter calculations have been made for known parameter distributions. A distinction can be made between the a priori distribution of $\underline{\theta}$ which is assumed before any measurements are made and the actual distribution of $\underline{\theta}$, which is present in the investigated measurements. An estimator can only be based on the a priori distribution because that is the only one that is known. The error that will be made by using such an estimator depends on the actual distribution. If these distributions differ the estimator is not optimal and peaking becomes possible.

The question may be raised whether there are practical problems in which the parameter distribution is really known. In such a problem the a priori density equals the actual density which makes it possible to construct the optimal estimators. It will be clear that these estimators are only optimal if one really is engaged with the complete class of classification problems. This is not the usual case in practice.

An example might be the following. Certain types of heart defects may cause for one subject two different types of electrocardiograms. One corresponds with normal behaviour of the heart and one with abnormal behaviour. For clinical purposes it is relevant to observe how often each of the two types are present in the electrocardiogram. This causes a classification problem. A learning set can be found by classifying a part of the cardiograms by man, after that the other cardiograms may be recognized automatically.

As the shape of both the normal as well as the abnormal cardiogram, may be subject dependent, one has for each subject a different classification problem. So there is a class of problems if more subjects are considered. From a number of subjects the distribution of the parameters $\underline{\theta}$ over this class

of problems may be estimated. Using this parameter distribution an optimal estimator can be calculated for the parameter vector of a new subject.

If the variance of the distribution of $\underline{\theta}$ is small, the cardiograms of different subjects are similar and only a small learning set is necessary for estimating the parameters $\underline{\theta}$ of a new subject. For a widely spread distribution of $\underline{\theta}$ more learning objects will be necessary. For a very widely spread distribution so many learning objects will be necessary that it hardly pays to use this a priori distribution. In that case the cardiograms of different subjects differ so much that the knowledge of the cardiograms of other subjects is of no use for recognizing the two types of cardiograms of a new subject.

In contrast with the above example often the parameter distribution is unknown. For those cases the relevance of this chapter is that it shows how the classification error can be studied given certain estimators and parameter distributions. This may give some arguments for the selection of the estimators and the feature size in relation to the sample size.

Chapter 5
MODEL COMPLEXITY

In the previous chapters the influences of feature size and sample size on the classification error have been considered. The density functions $f_A(\underline{x}|\underline{\theta}_A)$ and $f_B(\underline{x}|\underline{\theta}_B)$ were assumed to be known except for the parameters $\underline{\theta}_A$ and $\underline{\theta}_B$. In many practical situations, however, even the functional form of $f_\ell(\underline{x}|\underline{\theta}_\ell)$ is unknown. We will call a choice for that functional form the *statistical model*. In this chapter some effects of choosing a wrong statistical model will be considered. By this a family of density functions is meant that does not enclose the density of the population in question. Special attention will be given to the effect of the *model complexity*, i.e. the number of parameters involved in a certain model. The way an additional parameter influences the classification error is very complicated and differs from problem to problem. We will restrict ourselves, therefore, to some simple examples using artificially generated data.

First it is shown that the choice of a wrong model does not necessarily cause worse results. Models with a lower model complexity than the true one may result in a lower classification error. This result has, strictly spoken, already been shown in the previous chapter when we considered the peaking effect. In that case the use of all relevant features lead to worse results compared with the use of only a few of them. It will be shown here that such a result is also true for the covariances between the features. Neglecting covariances between correlated features may cause better classification results.

In a second example the last mentioned result will be illustrated, especially in relation to the value of the correlation coefficients. It appears in that example that the higher the correlation coefficients in the true model, the less likely that neglection of those correlation coefficients improves the results. Assuming zero correlation between lightly correlated features may in that case be better than estimating the correlation coefficients using a small sample set. The important fact for practice, therefore, is that it is

not so relevant how sure we are about the existence of some correlation between the features, but far more how large that correlation may be.

It will now be shown that there exists a model with specified parameters $\underline{\theta}$ such that the expected classification error $\bar{\varepsilon}$ decreases if the model is simplified, in spite of the fact that the more complex model is the correct one. As an example will be used two two-dimensional normal distributions (see fig. 5.1) with expectations

$$\underline{\mu}_A = (0, 0)$$
$$\underline{\mu}_B = (2,-1) \tag{5.1}$$

and covariance matrices

$$\Sigma_A = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$$
$$\Sigma_B = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix} \tag{5.2}$$

This implies that for class A the features are uncorrelated and that for class B the correlation coefficient is 0.5. Equal a priori probabilities will be assumed, so c = 0.5. The following four discriminant functions, in decreasing order of model complexity are used:

a) The Bayes discriminant function

$$R(\underline{x}) = \tfrac{1}{2}(\underline{x}-\underline{\mu}_B)^T \Sigma_B^{-1} (\underline{x}-\underline{\mu}_B) - \tfrac{1}{2}(\underline{x}-\underline{\mu}_A)^T \Sigma_A^{-1} (\underline{x}-\underline{\mu}_A) + \tfrac{1}{2} \ln\{\frac{|\Sigma_B|}{|\Sigma_A|}\} \tag{5.3}$$

The model complexity for this case will be coded as 4. This quadratic function follows immediately from (1.3) after substitution of the normal density function, see Fukunaga [23].

b) The linear discriminant function

$$R(\underline{x}) = (\underline{\mu}_A-\underline{\mu}_B)^T \Sigma^{-1} \underline{x} - \tfrac{1}{2} \underline{\mu}_A^T \Sigma^{-1} \underline{\mu}_A + \tfrac{1}{2} \underline{\mu}_B^T \Sigma^{-1} \underline{\mu}_B \tag{5.4}$$

in which $\Sigma = \frac{1}{2}(\Sigma_A + \Sigma_B)$. The model complexity for this case is coded as 3. The discriminant functions (5.4) and (5.3) are identical if $\Sigma_A = \Sigma_B$. The function

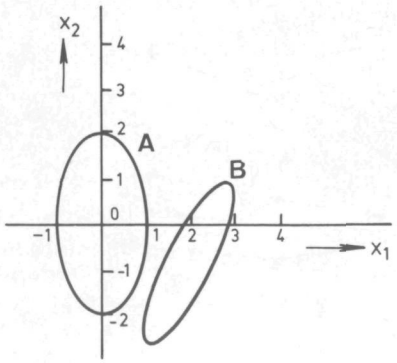


Fig. 5.1 The two classes as given by (5.1) and (5.2).

(5.4) is often used as a linear approximation of (5.3).

c) The linear discriminant function

$$R(\underline{x}) = (\underline{\mu}_A - \underline{\mu}_B)^T \Lambda^{-1} \underline{x} - \tfrac{1}{2} \underline{\mu}_A^T \Lambda^{-1} \underline{\mu}_A + \tfrac{1}{2} \underline{\mu}_B^T \Lambda^{-1} \underline{\mu}_B \qquad (5.5)$$

in which Λ is a diagonal matrix. The model complexity for this case is coded as 2. The discriminant functions (5.3), (5.4) and (5.5) are identical if Σ_A and Σ_B are identical and equal to Λ.

d) The linear discriminant function

$$R(\underline{x}) = (\underline{\mu}_A - \underline{\mu}_B) \cdot \underline{x} - \tfrac{1}{2} \underline{\mu}_A \cdot \underline{\mu}_A + \tfrac{1}{2} \underline{\mu}_B \cdot \underline{\mu}_B \qquad (5.6)$$

The model complexity for this case is coded as 1. The four discriminant functions are identical if the two covariance matrices are identical and diagonal and all variances are equal (of course this does not apply for the estimates of $R(\underline{x})$).

From each of the two classes with parameters as defined by (5.1) and (5.2) m learning objects were chosen at random. $R(\underline{x})$ was estimated using the plug-in rule and maximum likelihood estimates for Σ_A, Σ_B, Σ, Λ, $\underline{\mu}_A$ and $\underline{\mu}_B$. The classification error of each of the four discriminant functions was estimated by applying them to 200 test objects. This test set was the same for each discriminant function. This was repeated ten times for different learning

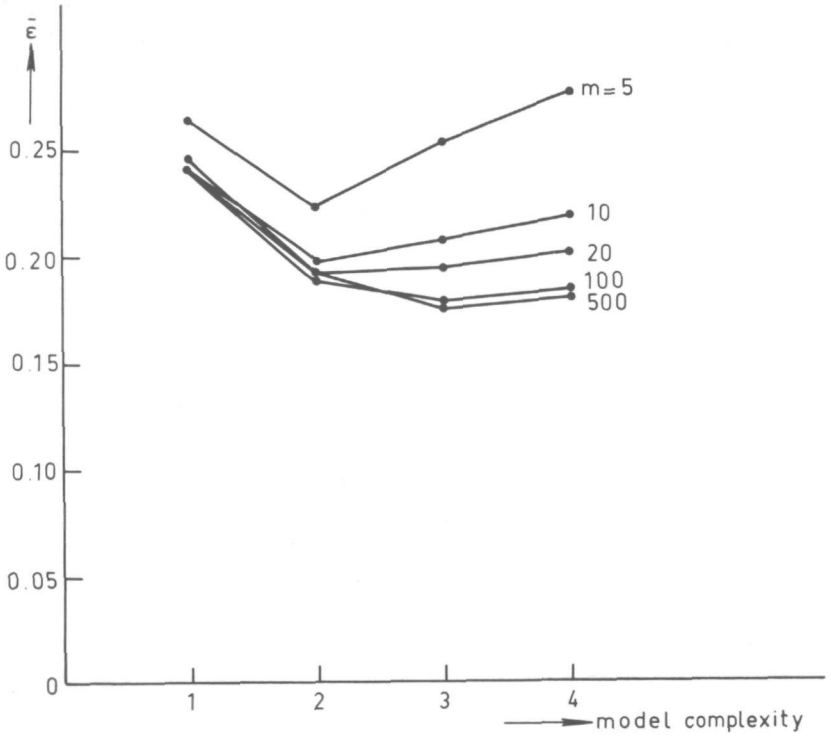


Fig. 5.2 The expected classification error $\bar{\varepsilon}$ as a function of model complexity and sample size for a two dimensional example. Only the measurement points have meaning. The lines are drawn just for clarity.

m	model complexity			
	1	2	3	4
5	0.265(0.023)	0.223(0.019)	0.254(0.024)	0.277(0.029)
10	0.242(0.019)	0.198(0.008)	0.209(0.007)	0.220(0.009)
20	0.246(0.014)	0.193(0.005)	0.195(0.004)	0.202(0.006)
100	0.242(0.003)	0.189(0.002)	0.178(0.003)	0.185(0.003)
500	0.241(0.003)	0.193(0.001)	0.176(0.002)	0.181(0.001)

Table 5.1 The expected classification error $\bar{\varepsilon}$ of a two-dimensional example as a function of sample size and model complexity (see text). The given numbers are the results of a Monte Carlo procedure and the computed standard deivations in those results.

sets. The averaged results are an estimate for $\bar{\varepsilon}$. They are presented in fig. 5.2; see also table 5.1 where in addition the computed standard deviations in the averages are given.

It appears that a peaking phenomenon occurs. A higher model complexity may result in worse performances. The optimal model complexity seems to increase for increasing sample size m. A similar effect for binary features has been shown by Schinkel [37].

The observed peaking phenomenon can be elucidated by realising that if the true value of a parameter is known to be small it may be estimated better by putting it equal to zero than by using a small learning set. Besides, a more simple discriminant function is obtained. Similarly, if two parameters are known to be almost equal it may be better to assume equality than to estimate two different parameters. The criterion for a better estimate is here the classification error.

The above mentioned importance of the correlation coefficient will be illustrated by a two-dimensional example in which the covariance matrices are chosen to be equal,

$$\Sigma = \Sigma_A = \Sigma_B = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \tag{5.7}$$

and in which the means are given by

$$\begin{aligned} \underline{\mu}_A &= (0, 0) \\ \underline{\mu}_B &= (1, 0) \end{aligned} \tag{5.8}$$

The expected classification error $\bar{\varepsilon}$ is estimated using a Monte Carlo procedure, in which the discriminant functions given by (5.4) and (5.5) were computed for 200 randomly generated learning sets. For each discriminant function and each learning set ε was computed analytically. The results, averaged over all learning sets are an estimate for $\bar{\varepsilon}$. In fig. 5.3a and fig. 5.3b $\bar{\varepsilon}$ is given for two values of m as a function ρ. Equal results are assumed for negative and positive values of ρ. See also table 5.2a and table 5.2b for the standard deviations. It appears that for small values of ρ it is better to assume $\rho = 0$ than to estimate ρ. The regions of ρ-values for which this is true shrinks with increasing m.

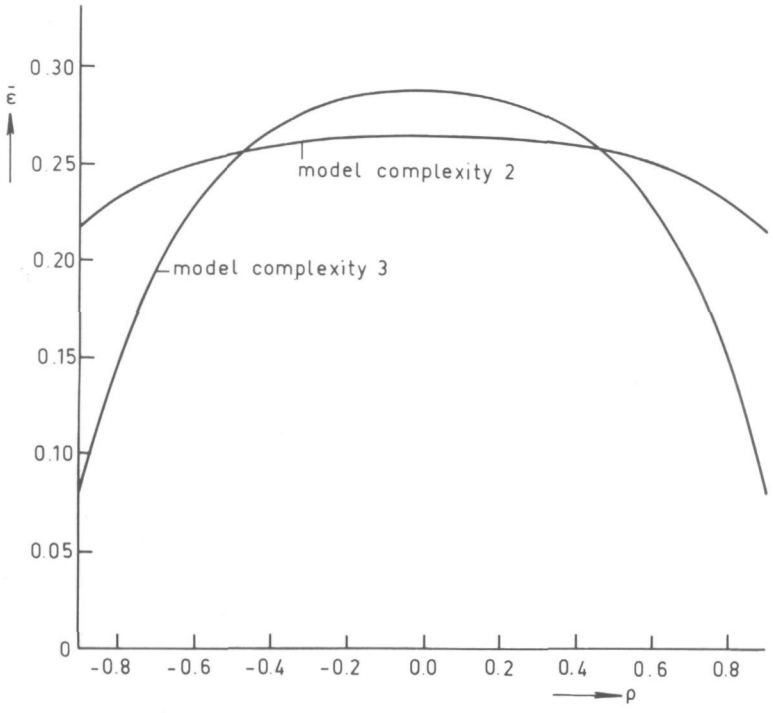


Fig. 5.3.a The expected classification error $\bar{\varepsilon}$ as a function of ρ for the presented example, m = 2.

ρ	model complexity	
	2	3
0.0	0.264(0.008)	0.290(0.009)
0.1	0.264(0.008)	0.286(0.009)
0.3	0.262(0.008)	0.278(0.009)
0.5	0.256(0.008)	0.253(0.009)
0.7	0.244(0.008)	0.197(0.009)
0.9	0.216(0.008)	0.081(0.007)

Table 5.2.a The expected classification error $\bar{\varepsilon}$ as a function of ρ for the presented example, m = 2.

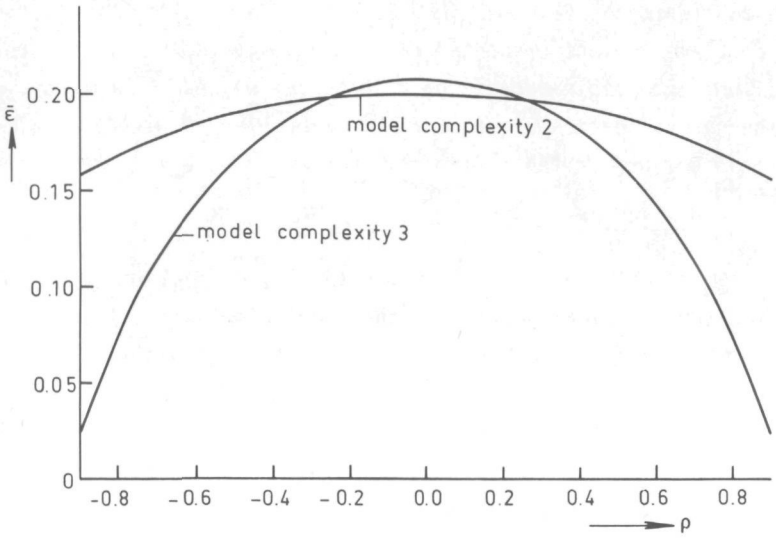


Fig. 5.3.b The expected classification error $\bar{\varepsilon}$ as a function of ρ for the presented example, m = 5.

ρ	model complexity	
	2	3
0.0	0.200(0.004)	0.209(0.004)
0.1	0.200(0.004)	0.207(0.004)
0.3	0.197(0.003)	0.195(0.003)
0.5	0.190(0.003)	0.167(0.003)
0.7	0.177(0.004)	0.116(0.003)
0.9	0.158(0.005)	0.025(0.002)

Table 5.2.b The expected classification error $\bar{\varepsilon}$ as a function of ρ for the presented example, m = 5.

The conclusion of this chapter is that the expected classification error $\bar{\varepsilon}$ may behave for increasing model complexity in a similar way as for increasing measurement complexity or feature size. Quantitative results which hold for a class of problems are not presented here. Analytical solutions are very difficult if not impossible because of the complex way additional parameters influence the classification error. For the same reason Monte Carlo simulations are impractical because of the difficulty in getting results with general validity.

Finally a remark will be made on the possibility of peaking of the mean classification error $\tilde{\varepsilon}$ as a function of the model complexity. The results of the previous chapter may apply here too. If for the estimators of the parameters the Bayes estimators are chosen based on the true parameter distribution then the discriminant function is optimal. This implies that if the model is such that the parameter space of a complex model encloses the parameter space of a more simple model then $\tilde{\varepsilon}$ will be minimum for the complex model and peaking does not occur.

Chapter 6
A PRIORI KNOWLEDGE

In this chapter the foregoing discussions will be considered from the point of view of a priori knowledge. Most points are already implicitly or explicitly mentioned. They are put together here in order to illustrate the importance of a priori knowledge.

If in a particular classification problem the sample size is as large as possible in the practical situation and if the features, the statistical model and the estimators are chosen optimally as far as knowledge reaches and if the classification error is still unsatisfactorily large, the only thing one can try to do is to enlarge the a priori knowledge. Especially knowledge of the following kinds will be useful.

- A priori knowledge on the discriminating power of the features. From the chapters 3 and 4 it follows that it is important for the accuracy and for avoiding peaking to keep the number of features as low as possible. The features have therefore to be ranked in such a way that ε^* is minimum for each number of features because by this the estimation error is kept as small as possible by constant ε^*.
- A priori knowledge on the relations between the features can improve the feature ranking. Sometimes one of two good features can be deleted because of their dependency.
- A priori knowledge on the class densities. This may result in the choice of better estimators as shown in 4.4. A decision has to be made whether or not certain parameters are taken into account in relation to the sample size, see chapter 5.
- A priori knowledge on the distribution of the parameters. This is applicable if the same classification system has to be used several times for a class of problems, see 4.5.

The estimators, the feature size and the model may be chosen in a way described in the previous chapters on the basis of the sample size and the

knowledge as described above. If the result is still not satisfactory the final thing one can do is search for new and better features. A priori knowledge in the context of statistical pattern recognition can be seen, as appears from the above, as any knowledge on the class distributions of possible features that is useful for selecting the discriminant procedure.

It might seem possible to build up a part of the a priori knowledge by making use of the learning set. This may be a trap, however, if the same learning objects are used for estimating the discriminant function. In that case no real a priori knowledge is used, only a more complicated discriminant procedure has been followed. For such a procedure the limitations discussed in the previous chapters are valid. Thus, real a priori knowledge has to be built up by other objects than the ones used for the discriminant analysis.

At this point we encounter a general problem. Discriminant procedures and other statistical techniques for gaining knowledge may be automized, for instance by using a computer. The question may be raised whether all a priori knowledge can be found by an automatic analysis of learning objects or an additional source of knowledge is needed. If the first is true it would be possible for very large learning sets to construct some discriminant procedure that does not make use of any a priori knowledge at all. If the second is true man is more than just a very complicated computer (see Turing [42]). This question and related problems on "computer knowledge" are discussed by Popper in "Objective knowledge" [34].

In this thesis an argument can be found for the impossibility of the first alternative. In chapter 4 has been shown that only under restricted conditions (e.g. having the right a priori knowledge) peaking can be avoided. In general, even for very large, but finite sample sizes the classification error shows peaking. This implies that if the feature size is large enough, discrimination becomes worse. If no a priori knowledge is given at all, anything can be a feature and the feature size is very large, if not infinite. Peaking is therefore to be expected in the absence of a priori knowledge. As many practical discriminant procedures give good results a source of knowledge has to exist that differs essentially from the analysis of observations as done in statistical pattern recognition.

In epistemology roughly three sources of knowledge are distinguished. The first is the observation, which was strongly emphasized by the empirists. In the above is argued that an automatic analysis of observations seems to be

insufficient for gaining knowledge (in our restricted sense of ability to discriminate). The second is the mind, as emphasized by the rationalists. The third is a combination, e.g. the process of conjecture and refutation as defined by Popper [32], or the role of thought as a bridge between intuition and observation see Steiner [40]. As the peaking phenomenon argues against pure observation as a source of knowledge, even in combination with a "computer mind", it argues for the existence of something else, such as intuition.

Chapter 7
CONCLUSIONS AND DISCUSSION

The classification error ε of a statistical pattern recognizer is expressed by an upper bound into the Bayes error and the estimation errors of the class distributions. The expectation of these estimation errors can be computed for certain distributions like an arbitrary normal distribution, as a function of sample size and feature size. This results in an upper bound of the expected classification error $\bar{\varepsilon}$ expressed in sample size and feature size. It could also be shown, however, that better estimates, thus smaller estimation errors, do not necessarily yield a smaller classification error (chapter 3).

The Bayes error decreases with increasing feature size, but the expected estimation error increases. The expected classification error, therefore, may increase or decrease. If it increases this is called peaking. The causes of the peaking phenomenon are the choice of bad estimators and a too large variance of the estimated discriminant function compared with the small contribution of the new feature to the discriminating power (chapter 4).

It could be shown, however, that the mean classification error, which is the expected classification error averaged over a class of problems, shows no peaking if Bayes estimators are used, based on the parameter distribution of that class of problems (§ 4.4). An explanation of the result that by use of this estimator in a model presented by Hughes [24] peaking still occurs with increasing measurement complexity (§ 4.2) appeared to be the fact that increasing measurement complexity is virtually not the same as the addition of a new feature (§ 4.4).

Conditions were presented for a class of distributions and a class of estimators for which the mean classification error does not peak (§ 4.3).

In the case of increasing model complexity, in which more parameters, describing the statistical model, are added, instead of features, a similar peaking phenomenon can be observed. This was shown by an example (chapter 5).

If enough a priori knowledge is available it is possible to compute along the presented lines whether peaking is to be expected in a given situation.

An important problem which has not been investigated in this thesis is how peaking can be detected in a given classification problem. A very short discussion follows. If the sample size is large the learning set can be split into a part for learning and a part for testing. Peaking may be detected using the test set. In case peaking occurs, however, it is certainly better to use all objects for learning and no objects are left for testing. A number of methods is available for the estimation of the classification error with an economic use of test objects, see for instance Toussaint [41] and Lissack and Fu [28]. These estimators differ in variance and biasedness. For the detection of peaking in a single case, using a single learning set, however, unbiasedness is not essential. An estimator with a small variance might be better if its mean square error is smaller. It will still be possible, however, that peaking is detected too late, or not at all. For that reason an investigation of the circumstances under which peaking might happen as has been presented here may be useful.

Before estimating a discriminant function using a small learning set, one should realize that the result can never be better than the learning set permits. A too detailed method will determine the discriminant function on some incidental details of the learning set which are not representative for the population. The larger the learning set, the more reliable its details and the more significant the use of a detailed method.

For a scientist, for instance a physicist, who wants to use a method like statistical pattern recognition for reaching his conclusions, this thesis implies that he has to limit the number of variables he wants to measure. Such a method might be of use if it is difficult to obtain sufficient accurate physical model, for instance in the case of weather-forecasting. Under these circumstances often a statistical approach is chosen. This implies that many observations have to be available. The variables, like temperature, air pressure and humidity and the measurement places have to be chosen on the basis of a priori knowledge and their number should be limited. If that number is large, it is hard to draw significant conclusions on statistical grounds. Now use has to be made of a physical model for relating the variables. After this has been done statistical pattern recognition may not be necessary anymore or the statistical problem is simplified and conclusions are more

significant. It is not possible to construct such a physical model completely on the ground of the statistics of the learning set, because in that case a similar inaccuracy is introduced as by using statistical models. It is necessary to use physical knowledge and intuition for the construction of the model from the statistical data. The elements: observations, a priori knowledge, thought and intuition constitute together the basis for new knowledge. Statistical pattern recognition can be an aid for that.

APPENDICES

APPENDIX A

THE "WORST" PROBABILITY FUNCTION IN THE GENERAL MEASUREMENT SPACE.

Here it will be proved that (in the case of even n) $\frac{1}{2}E_\chi(e_A) + \frac{1}{2}E_\chi(e_B)$, with $E_\chi(e_\ell)$ given by (3.33)

$$\tfrac{1}{2}E_\chi(e_A) + \tfrac{1}{2}E_\chi(e_B) = \tfrac{1}{2}\sum_{j=1}^{n}\{p_A^j(1-p_A^j)/(2\pi m)\}^{\frac{1}{2}} + \tfrac{1}{2}\sum_{j=1}^{n}\{p_B^j(1-p_B^j)/(2\pi m)\}^{\frac{1}{2}} \qquad (A.1)$$

is maximum if

$$p_A^j = 2\varepsilon^*/n \qquad (A.2)$$

$$p_B^j = 2(1-\varepsilon^*)/n \qquad (A.3)$$

for n/2 values of j (n even), and if

$$p_A^j = 2(1-\varepsilon^*)/n \qquad (A.4)$$

$$p_B^j = 2\varepsilon^*/n \qquad (A.5)$$

for the other n/2 values of j. Constraints for the maximization are

$$\sum_{j=1}^{n} p_\ell^j = 1 \qquad (\ell = A,B) \qquad (A.6)$$

$$\varepsilon^* = \tfrac{1}{2}\sum_{j=1}^{n} \min\{p_A^j, p_B^j\} \qquad (A.7)$$

Let $J = \{j_1, j_2, \ldots j_q\}$ be a subset of the set of indices $\{0, 1, 2, \ldots n\}$ and let $p_A^j < p_B^j$ for $j \in J$ and $p_A^j \geq p_B^j$ for $j \notin J$. Condition (A.7) becomes in that case

$$\varepsilon^* = \tfrac{1}{2} \sum_{j\in J} p_A^j + \tfrac{1}{2} \sum_{j\notin J} p_B^j \qquad (A.8)$$

It is easily seen that $q \neq 0$ and $q \neq n$ because in those cases (A.8) would give $\varepsilon^* = \frac{1}{2}$, and can therefore not be chosen arbitrarily.

Now the maximum of (A.1) may be found by using the Lagrange multiplier method. The function to be maximized is, apart from the term $(2\pi m)^{-\frac{1}{2}}$ in (A.1),

$$U = \tfrac{1}{2} \sum_{j=1}^{n} \{p_A^j(1-p_A^j)\}^{\frac{1}{2}} + \tfrac{1}{2} \sum_{j=1}^{n} \{p_B^j(1-p_B^j)\}^{\frac{1}{2}} +$$
$$\lambda_1 \{\sum_{j=1}^{n} p_A^j - 1\} + \lambda_2 \{\sum_{j=1}^{n} p_B^j - 1\} + \lambda_3 \{\tfrac{1}{2} \sum_{j=1}^{n} \min\{p_A^j, p_B^j\} - \varepsilon^*\} \qquad (A.9)$$

The derivative of U to p_A^j is

$$\tfrac{1}{4}\{p_A^j(1-p_A^j)\}^{-\frac{1}{2}} (1-2p_A^j) + \lambda_1 + \tfrac{1}{2}\lambda_3 \qquad \text{for } j\in J \qquad (A.10)$$

$$\tfrac{1}{4}\{p_A^j(1-p_A^j)\}^{-\frac{1}{2}} (1-2p_A^j) + \lambda_1 \qquad \text{for } j\notin J \qquad (A.11)$$

If these derivatives are put equal to zero it appears that the optimal p_A^j is a constant for $j\in J$. The same applies for $j\notin J$, so $p_A^j = p_A$ for $j\in J$ and $p_A^j = p_A'$ for $j\notin J$. Analogously from the derivatives of U to p_B^j

$$\tfrac{1}{4}\{p_B^j(1-p_B^j)\}^{-\frac{1}{2}} (1-2p_B^j) + \lambda_2 \qquad \text{for } j\in J \qquad (A.12)$$

$$\tfrac{1}{4}\{p_B^j(1-p_B^j)\}^{-\frac{1}{2}} (1-2p_B^j) + \lambda_2 + \tfrac{1}{2}\lambda_3 \qquad \text{for } j\notin J \qquad (A.13)$$

it follows that $p_B^j = p_B$ for $j\in J$ and $p_B^j = p_B'$ for $j\notin J$. Using these results the constraints (A.6) and (A.7) can be written as

$$qp_A + (n-q)p_A' = 1 \qquad (A.14)$$

$$qp_B + (n-q)p_B' = 1 \qquad (A.15)$$

$$\varepsilon^* = \tfrac{1}{2}qp_A + \tfrac{1}{2}(n-q)p_B' \qquad (A.16)$$

With (A.14) p_A' can be expressed in p_A, with (A.16) p_B' can be expressed in p_A

and with this result and (A.15) it is possible to express p_B in p_A.

$$p'_A = (1 - qp_A)/(n - q) \tag{A.17}$$

$$p'_B = 2(\varepsilon^* - \tfrac{1}{2} qp_A)/(n - q) \tag{A.18}$$

$$p_B = (1 - \tfrac{1}{2}\varepsilon^* + qp_A)/q \tag{A.19}$$

If the derivatives (A.10) - (A.13) are put equal to zero and λ_1, λ_2 and λ_3 are eliminated, one obtains

$$\{p_A(1-p_A)\}^{-\frac{1}{2}}(1-2p_A) - \{p'_A(1-p'_A)\}^{-\frac{1}{2}}(1-2p'_A)$$

$$- \{p'_B(1-p'_B)\}^{-\frac{1}{2}}(1-2p'_B) + \{p_B(1-p_B)\}^{-\frac{1}{2}}(1-2p_B) = 0 \tag{A.20}$$

All the four terms of (A.20) are monotonous decreasing functions of p_A because p'_A and p'_B are linear decreasing functions of p_A and p_B is a linear increasing function of p_A, as follows from (A.17) - (A.19). This implies that (A.20) has at most one solution for p_A. This solutions yields a maximum for U because the derivatives (A.10) - (A.13) are monotonous decreasing functions of p^j_A and p^j_B. For U can now be written

$$U = \tfrac{1}{2} q \{p_A(1-p_A)\}^{\frac{1}{2}} + \tfrac{1}{2}(n-q) \{p'_A(1-p'_A)\}^{\frac{1}{2}}$$

$$+ \tfrac{1}{2} q \{p_B(1-p_B)\}^{\frac{1}{2}} + \tfrac{1}{2}(n-q) \{p'_B(1-p'_B)\}^{\frac{1}{2}} \tag{A.21}$$

This is a function of p_A and q after substitution of (A.17) - (A.19).

$$U = \tfrac{1}{2}\{p_A q(q-p_A q)\}^{\frac{1}{2}} + \tfrac{1}{2}\{(1-p_A q)(n-q-1+p_A q)\}^{\frac{1}{2}}$$

$$+ \tfrac{1}{2}\{(1-2\varepsilon^*+p_A q)(q-1+2\varepsilon^*-p_A q)\}^{\frac{1}{2}}$$

$$+ \tfrac{1}{2}\{(2\varepsilon^*-p_A q)(n-q-2\varepsilon^*+p_A q)\}^{\frac{1}{2}} \tag{A.22}$$

For the following q will be considered as a continuous variable between 0 and n. The derivative to q for constant p_A is given by

$$\frac{\partial U}{\partial q} = \tfrac{1}{4}\{p_A q(q-p_A q)\}^{-\frac{1}{2}}\,(2p_A q - 2p_A^2 q)$$

$$+\ \tfrac{1}{4}\{(1-p_A q)(n-q-1+p_A q)\}^{-\frac{1}{2}}\,(-1+p_A-p_A n+2p_A q+p-2p_A^2 q)$$

$$+\ \tfrac{1}{4}\{(1-2\varepsilon^*+p_A q)(q-1+2\varepsilon^*-p_A q)\}^{-\frac{1}{2}}(1-p_A-2\varepsilon^*+2\varepsilon^* p_A+2p_A q+p\ +2p_A\varepsilon^*-2p_A^2 q)$$

$$+\ \tfrac{1}{4}\{(2\varepsilon^*-p_A q)(n-q-2\varepsilon^*+p_A q)\}^{-\frac{1}{2}}\,(-2\varepsilon^*+2\varepsilon^* p_A-p_A n+2p_A q+2p_A\varepsilon^*-2p_A^2 q) \qquad (A.23)$$

The probabilities p'_A, p_B and p'_B as given by (A.17) - (A.19) have values between 0 and 1 only if

$$0 < q \leqslant 2\varepsilon^*/p_A \qquad (A.24)$$

It can be verified that the four terms of (A.23) are monotonous non-increasing functions of q for the interval given by (A.24). This implies that the zero crossing of (A.23) on that interval gives the absolute maximum of U as a function of q. The proof is completed by verifying that the solution as given by (A.2) - (A.5), which implies that $p_A = 2\varepsilon^*/n$ and $q = n/2$, satisfies (A.20) and makes (A.23) equal to zero.

APPENDIX B

LEAST SQUARES APPROXIMATION AND THE PEAKING PHENOMENON

In this appendix it will be shown that the error of a least squares approximation shows a similar kind of peaking as is described in chapter 4 for the classification error.

Let the function $y = F(\underline{x})$ be observed for the values $\underline{x}_i$ $(i = 1,m)$ of the vector $\underline{x}$. Denote the observations by y_i. An approximation of y will be made using a set of not necessarily orthogonal basis function $\varphi_j(\underline{x})$ $(j = 1,k;$ $k < m)$.

$$\hat{y} = \sum_{j=1}^{k} a_j \varphi_j(\underline{x}) \tag{B.1}$$

Let $\hat{y}_i$ be the value of $\hat{y}$ for $\underline{x} = \underline{x}_i$. The coefficients a_j are chosen such that

$$\hat{\delta} = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{B.2}$$

is minimum. In vector notation this can be written as

$$\hat{\delta} = (\underline{y} - \hat{\underline{y}})^T (\underline{y} - \hat{\underline{y}}) \tag{B.3}$$

Substitution of (B.1) gives

$$\hat{\delta} = (\underline{y} - \varphi^T \underline{a})^T (\underline{y} - \varphi^T \underline{a}) \tag{B.4}$$

in which

$$\underline{y} = (y_i;\ i = 1,m)^T \tag{B.5}$$

$$\underline{a} = (a\ ;\ j = 1,k)^T \tag{B.6}$$

and φ is a matrix given by

$$\varphi = (\varphi_j(\underline{x}_i);\ j = 1,k;\ i = 1,m) \tag{B.7}$$

(B.4) is minimum if

$$\underline{a} = (\varphi\varphi^T)^{-1}\ \varphi\ \underline{y} \tag{B.8}$$

These formula will be used for the case that $\underline{y}$ is random and that (B.1) should approximate the expectation of $\underline{y}$, which will be defined as

$$E(\underline{y}) = \underline{\mu} \tag{B.9}$$

Assume that the variances of all y_i are equal.

$$Var(y_i) = \sigma_r^2 \qquad (i = 1,m) \tag{B.10}$$

Assume that the values of all y_i's are uncorrelated

$$E(y_i y_j) = E(y_i)\ E(y_j) \qquad (\forall i \neq j) \tag{B.11}$$

For the expected square error $\overline{\delta}$, defined as $E\{(\underline{\mu}-\hat{\underline{y}})^T(\underline{\mu}-\hat{\underline{y}})\}$ can be written

$$\overline{\delta} = \frac{1}{m}\ E\{(\underline{\mu} - \varphi^T\underline{a})^T\ (\underline{\mu} - \varphi^T\underline{a})\} \tag{B.12}$$

After substitution of (B.8) one finds

$$\overline{\delta} = \frac{1}{m}\ E\{\underline{\mu}^T\underline{\mu} - 2\underline{\mu}^T\varphi^T(\varphi\varphi^T)^{-1}\ \varphi\underline{y} + \underline{y}^T\varphi^T(\varphi\varphi^T)^{-1}\ \varphi\underline{y}\} \tag{B.13}$$

in which use has been made of the fact that $\varphi\varphi^T$ is a symmetric matrix. Let the matrix D be given by

$$\mathrm{D} = \varphi^T(\varphi\varphi^T)^{-1}\ \varphi \tag{B.14}$$

The k rows of φ are eigenvectors of D because $\mathrm{D}\varphi^T = \varphi^T$. This set of eigenvectors is complete because $\mathrm{Rank}(\mathrm{D}) = k$. All corresponding eigenvalues are one. So the trace of D, which is equal to the sum of the eigenvalues,is k. For (B.13) can now be written after some calculations

$$\overline{\delta} = \frac{1}{m} (\underline{\mu}^T\underline{\mu} - \underline{\mu}^T D\underline{\mu} + tr(D)\sigma_r^2) \qquad (B.15)$$

in which use has been made of (B.9) - (B.11).

For (B.15) the following geometric interpretation can be given. The inner product $\underline{\mu}^T\underline{\mu}$ is the square distance of $\underline{\mu}$ to the origin. $D\underline{\mu}$ is the projection of $\underline{\mu}$ on the space R_k spanned by the eigenvectors of D (which are the rows of φ). $\underline{\mu}^T D\underline{\mu}$ is the square distance of $D\underline{\mu}$ to the origin because $D^TD = D$. So $\underline{\mu}^T\underline{\mu} - \underline{\mu}^T D\underline{\mu}$ is the square distance of $\underline{\mu}$ to R_k. This is a monotonous non-increasing function of k. The term $tr(D)\sigma_r^2 = k\sigma_r^2$ increases linearly with k. It is therefore possible that $\overline{\delta}$ peaks as a function of k, see fig. B.1. The expected error $\overline{\delta}$ decreases monotonically if σ_r^2 is small and increases monotonically if σ_r^2 is large. If $\underline{\mu}$ is completely described by the first n functions φ_j, then $\underline{\mu}^T\underline{\mu} = \underline{\mu}^T D\underline{\mu}$ for $k \geqslant n$ and therefore $\overline{\delta} = k\sigma_r^2$ for $k \geqslant n$.

It is interesting to investigate the mean square error $\tilde{\delta}$, defined as the expectation of $\overline{\delta}$ over a class of signals in analogy with the mean classification error. As an example will be treated the case of signals with constant power $m\sigma_s^2$ that can be described completely by n functions φ_j. Assume that $\underline{\mu}$ is uniformly distributed on a hypersphere in R_n with radius $m\sigma_s^2$. The mean square error can now be written as

$$\tilde{\delta} = \frac{1}{m} E(\underline{\mu}^T\underline{\mu} - \underline{\mu}^T D\underline{\mu}) + \frac{k}{m}\sigma_r^2 \qquad (B.16)$$

The expectation is equivalent to the expected square distance of a point of the hypersphere in R_n to a k-dimensional subspace containing the centre of the sphere. This expectation is

$$E(\underline{\mu}^T\underline{\mu} - \underline{\mu}^T D\underline{\mu}) = \frac{n-k}{n} m\sigma_s^2 \qquad (n \geqslant k) \qquad (B.17)$$

Substitution into (B.16) yields

$$\tilde{\delta} = \frac{n-k}{n}\sigma_s^2 + \frac{k}{m}\sigma_r^2 \qquad (n \geqslant k) \qquad (B.18)$$

This is the mean square error of the description of an arbitrary signal by k basis functions, using m observations, if the signal can be completely described by the first n basis functions. If $k \geqslant n$ the first term of (B.18)

has to be put equal to zero. This implies that (B.18) shows peaking if $n < m$, because in that case $\tilde{\delta}$ decreases for $k < n$ and increases for $k > n$, see fig. B.2. The well known sample theorem ($k \geq n$, $k < m$ so choose $m > n$) can be derived from (B.18).

The situation described in this appendix is much more simple than the one in pattern recognition because there the square error is hardly used as criterion. The relation between the probability of error and the least square error criterion is studied by Devijver [10].

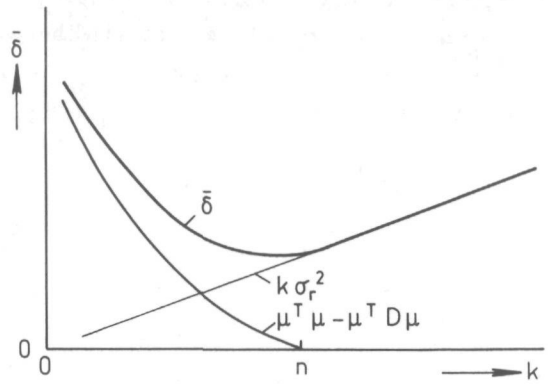


Fig. B.1 The expected square error $\bar{\delta}$ and its contributions (B.15).

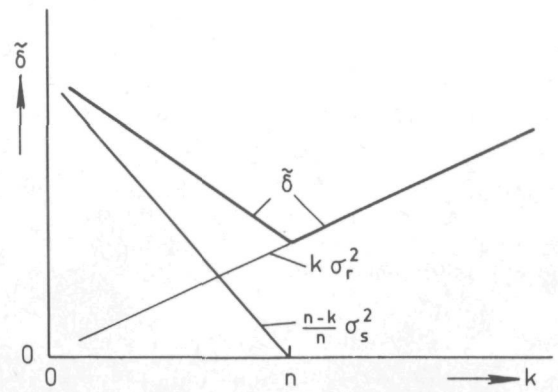


Fig. B.2 The mean square error $\tilde{\delta}$ and its contributions (B.18).

APPENDIX C

INDEPENDENT BINARY FEATURES AND THE PEAKING PHENOMENON

Peaking in the case of independent binary features is investigated here further. First the case of maximum likelihood estimates will be treated generally. In the second part of this appendix is shown for an example using Bayes estimators that the conditions for no peaking (4.17) and (4.18) are fulfilled.

The maximum likelihood estimates $\hat{p}_\ell^j$ ($j = 1,k$; $\ell = A,B$) are given by

$$\hat{p}_\ell^j = \frac{n_\ell^j}{m} \tag{C.1}$$

The discriminant function $\hat{S}(\underline{x})$ is found by substitution of

$$\hat{f}_\ell(\underline{x}) = \prod_{j=1}^{k} (\hat{p}_\ell^j)^{x_j} (1 - \hat{p}_\ell^j)^{1-x_j} \tag{C.2}$$

in (1.7)

$$\hat{S}(\underline{x}) = c\, \hat{f}_A(\underline{x}) - (1-c)\, \hat{f}_B(\underline{x}) \tag{C.3}$$

The probability that $\hat{p}_\ell^j = 0$ is given by

$$\text{Prob}(\hat{p}_\ell^j = 0) = (1 - p_\ell^j)^m \tag{C.4}$$

The probability that $\hat{p}_\ell^j = 1$ is given by

$$\text{Prob}(\hat{p}_\ell^j = 1) = (p_\ell^j)^m \tag{C.5}$$

For the probability that $\hat{S}(\underline{x}) = 0$ for an arbitrary learning set in an arbitrary point $\underline{x}$ can be written

$$\text{Prob}(\hat{S}(\underline{x}) = 0) \geqslant \text{Prob}(\hat{f}_A(\underline{x}) = 0 \text{ and } \hat{f}_B(\underline{x}) = 0) \tag{C.6}$$

It will be shown that, for finite m, the right part of (C.6) approaches one under general conditions if $k \to \infty$.

$$\text{Prob}(\hat{f}_A(\underline{x}) = 0 \text{ and } \hat{f}_B(\underline{x}) = 0) =$$

$$= 1 - \text{Prob}(\hat{f}_A(\underline{x}) \neq 0 \text{ or } \hat{f}_B(\underline{x}) \neq 0) \tag{C.7}$$

$$= 1 - \prod_{j=1}^{k} [\text{Prob}(x_j=1)\{1 - \text{Prob}(\hat{p}_A^j = 0)\ \text{Prob}(\hat{p}_B^j = 0)\} +$$

$$+ \text{Prob}(x_j=0)\{1 - \text{Prob}(\hat{p}_A^j = 1)\ \text{Prob}(\hat{p}_B^j = 1)\}] \tag{C.8}$$

The expression behind the product symbol in (C.8) equals one in the following cases:

1. $p_A^j = p_B^j = 1$. In this case is $\text{Prob}(\hat{p}_A^j = 0) = \text{Prob}(\hat{p}_B^j = 0) = 0$, see (C.4), and $\text{Prob}(\hat{p}_A^j = 1) = \text{Prob}(\hat{p}_B^j = 1) = 1$, see (C.5). Behind the product sign is therefore obtained $\text{Prob}(x_j = 1)$, which equals one.
2. $p_A^j = p_B^j = 0$. In this case is $\text{Prob}(\hat{p}_A^j = 0) = \text{Prob}(\hat{p}_B^j = 0) = 1$, see (C.4), and $\text{Prob}(\hat{p}_A^j = 1) = \text{Prob}(\hat{p}_B^j = 1) = 0$, see (C.5). Behind the product sign is therefore obtained $\text{Prob}(x_j = 0)$, which equals one.
3. $p_A^j = 1$, $p_B^j = 0$. In this case is $\text{Prob}(\hat{p}_A^j = 0) = 0$, see (C.4), and $\text{Prob}(\hat{p}_B^j = 1) = 0$, see (C.5). Behind the product sign is therefore obtained $\text{Prob}(x_j = 1) + \text{Prob}(x_j = 0)$, which equals one.
4. $p_A^j = 0$, $p_B^j = 1$. In this case is $\text{Prob}(\hat{p}_B^j = 0) = 0$, see (C.4), and $\text{Prob}(\hat{p}_A^j = 1) = 0$, see (C.5). Behind the product sign is therefore obtained $\text{Prob}(x_j = 1) + \text{Prob}(x_j = 0)$, which equals one.

In all other cases the expression behind the product sign is smaller than one. If the number of features for which this is the case increases then (C.8) approaches one provided that for all these features the expression mentioned is smaller than $1 - \delta$, with δ some arbitrary constant between zero and one. Under these conditions one finds

$$\lim_{k \to \infty} \text{Prob}(\hat{S}(\underline{x}) = 0) = 1 \tag{C.9}$$

If $\hat{S}(\underline{x}) = 0$ no discrimination is possible and the best thing one can do is

assign all points to the class with highest a priori probability, so

$$\bar{\varepsilon} = \min\{c, 1-c\} \tag{C.10}$$

For small k the classification error is usually smaller than min{c, 1-c}, which proves the existance of peaking in the case of maximum likelihood estimates.

The mean classification error $\tilde{\varepsilon}$ is the expectation of $\bar{\varepsilon}$ over some distribution of p_A^j and p_B^j. If the sum of the probabilities of the four cases mentioned above is smaller than one then also $\tilde{\varepsilon}$ shows peaking because (nearby) each $\bar{\varepsilon}$ peaks.

Now the possibility of peaking of $\tilde{\varepsilon}$ will be treated for the case of Bayes estimates using a uniform a priori density for the parameters. The following density estimate is found using (2.20)

$$\hat{f}^j(\underline{x}) = \prod_{j=1}^{k} \left\{ \left(\frac{n_\ell^j + 1}{m + 2} \right)^{x_j} \left(1 - \frac{n_\ell + 1}{m + 2} \right)^{1-x_j} \right\} \tag{C.11}$$

The discriminant function $\hat{R}(\underline{x})$ can be written as in (4.15)

$$R(\underline{x}) = \sum_{j=1}^{k} r^j - d \tag{C.12}$$

with r^j, see (4.42)

$$r^j = x^j \log\left(\frac{n_A^j+1}{n_B^j+1}\right) + (1-x^j) \log\left(\frac{m-n_A^j+1}{m-n_B^j+1}\right) \tag{C.13}$$

and $d = \log\{(1-c)/c\}$.

Now the proof will be given that (C.13) fulfils the condition for no peaking (4.17) if the joint density of p_A^j and p_B^j is uniform along the line $p_A^j = 1 - p_B^j$ and is zero elsewhere. In a similar way can be proved that condition (4.18) is also fulfilled. The indices j will be omitted. The expectation of r over $\underline{x} \in A$, χ and θ is written as $E_A(r)$.

$$E_A(r) = \sum_{n_A=0}^{m} \sum_{n_B=0}^{m} \int_{p_A} \int_{p_B} \left\{ p_A \log\left(\frac{n_A+1}{n_B+1}\right) + (1-p_A) \log\left(\frac{m-n_A+1}{m-n_B+1}\right) \right\}$$

$$\binom{m}{n_A} p_A^{n_A} (1-p_A)^{m-n_A} \binom{m}{n_B} p_B^{n_B} (1-p_B)^{m-n_B} \; h(p_A, p_B) \; dp_A \; dp_B \tag{C.14}$$

Introduce

$$U(n_A,n_B) = \log\left(\frac{n_A+1}{n_B+1}\right) \tag{C.15}$$

Note that

$$U(n_A,n_B) = - U(n_B,n_A) \tag{C.16}$$

and

$$U(n_A,n_B) > 0 \quad \text{for} \quad n_A > n_B \tag{C.17}$$

In (C.14) $p_B = 1 - p_A$ can be substituted on which line $h(p_A,p_B)$ is uniform.

$$E_A(r) = \sum_{n_A=0}^{m} \sum_{n_B=0}^{m} \int_{P_A} \left\{ p_A U(n_A,n_B) + (1-p_A)\ U(m-n_A+1,m-n_B+1) \right\}$$

$$\binom{m}{n_A}\binom{m}{n_B} p_A^{\,m+n_A-n_B}\ (1-p_A)^{\,m+n_B-n_A}\ dp_A \tag{C.18}$$

Define

$$W(n_A,n_B) = \binom{m}{n_A}\binom{m}{n_B} \int_{P_A} p_A^{\,m+n_A-n_B+1}\ (1-p_A)^{\,m+n_B-n_A}\ dp_A \tag{C.19}$$

The integral is a β-function, so

$$W(n_A,n_B) = \binom{m}{n_A}\binom{m}{n_B} \frac{(m+n_A-n_B+1)!\,(m-n_A+n_B)!}{(2m+2)!} \tag{C.20}$$

Note that

$$W(n_A,n_B) = \frac{m+n_A-n_B+1}{m+n_B-n_A}\ W(n_B,n_A) \tag{C.21}$$

Formula (C.18) can now be written as

$$E_A(r) = \sum_{n_A=0}^{m} \sum_{n_B=0}^{m} \left\{ U(n_A,n_B)\ W(n_A,n_B) + U(m-n_A,m-n_B)\ W(n_B,n_A) \right\} \tag{C.22}$$

This summation can be split into one over $n_A < n_B$ and one over $n_A > n_B$. The

term with $n_A = n_B$ is zero as $U(n_A,n_A) = 0$ because of (C.16). If n_A and n_B in the sum over $n_A < n_B$ are exchanged one finds

$$E_A(r) = \sum_{n_A>n_B} \Big\{ U(n_B,n_A)\, W(n_B,n_A) + U(m-n_B,m-n_A)\, W(n_A,n_B) + U(n_A,n_B)\, W(n_A,n_B) + U(m-n_A,m-n_B)\, W(n_B,n_A) \Big\} \tag{C.23}$$

which is equivalent to

$$E_A(r) = \sum_{n_A>n_B} \Big\{ U(n_A,n_B)\{- W(n_B,n_A) + W(n_A,n_B)\} + U(m-n_B,m-n_A)\{W(n_A,n_B) - W(n_B,n_A)\} \Big\} \tag{C.24}$$

because of (C.16).

From (C.21) follows

$$W(n_B,n_A) < W(n_A,n_B) \quad \text{for } n_A > n_B \tag{C.25}$$

Finally it can be concluded, using (C.17) and (C.25) that from (C.24) follows

$$E_A(r) > 0.$$

This completes the proof of (4.17) for the example under investigation.

APPENDIX D

THE ESTIMATION ERROR FOR NORMAL DENSITIES

In this appendix it will be shown that the expected estimation error for normal distributions, given by (3.39)

$$E_\chi(e) = 1 - E_\chi\left\{ \int_{\underline{x}} \min\{f(\underline{x}|\hat{\underline{\mu}},\hat{\Sigma}),\ f(\underline{x}|\underline{\mu},\Sigma)\}\ d\underline{x} \right\} \tag{D.1}$$

is independent of the expectation $\underline{\mu}$ and the covariance matrix Σ. The maximum likelihood estimates $\hat{\underline{\mu}}$ and $\hat{\Sigma}$ are the following functions of the set of learning objects $\chi = \{\underline{x}^1,\underline{x}^2,\ldots,\underline{x}^m\}$

$$\hat{\underline{\mu}} = \frac{1}{m}\sum_{i=1}^{m} \underline{x}^i \tag{D.2}$$

$$\hat{\Sigma} = \frac{1}{m}\sum_{i=1}^{m} \underline{x}^i\underline{x}^{iT} - \hat{\underline{\mu}}\hat{\underline{\mu}}^T \tag{D.3}$$

in which $\underline{x}^{iT}$ and $\hat{\underline{\mu}}^T$ are the transposes of the column vectors $\underline{x}^i$ and $\hat{\underline{\mu}}$. $E_\chi(e)$ can be written as

$$\begin{aligned} E_\chi(e) = 1 - \int_{\underline{x}^1}\int_{\underline{x}^2}\cdots\int_{\underline{x}^m}\int_{\underline{x}} \min\{&(2\pi)^{-k/2}|\hat{\Sigma}|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\underline{x}-\hat{\underline{\mu}})^T\hat{\Sigma}^{-1}(\underline{x}-\hat{\underline{\mu}})), \\ &(2\pi)^{-k/2}|\Sigma|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu}))\} \\ &\prod_{i=1}^{m}\{(2\pi)^{-k/2}|\Sigma|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\underline{x}^i-\underline{\mu})^T\Sigma^{-1}(\underline{x}^i-\underline{\mu}))\}\ d\underline{x}\ d\underline{x}^1\ d\underline{x}^2\ \ldots\ d\underline{x}^m \end{aligned} \tag{D.4}$$

The following substitutions will be made

$$\underline{x} = W\underline{z} + \underline{\mu} \tag{D.5}$$

$$\underline{x}^i = W\underline{z}^i + \underline{\mu} \qquad (i = 1,m) \tag{D.6}$$

in which W is defined by

$$\Sigma = WW^T \tag{D.7}$$

This is always possible because Σ is a positive definite square matrix. Note that

$$|\Sigma|^{\frac{1}{2}} = |W| \tag{D.8}$$

$$d\underline{x} = |W|\, d\underline{z} \tag{D.9}$$

$$d\underline{x}^i = |W|\, d\underline{z}^i \tag{D.10}$$

For $\hat{\underline{\mu}}$ can be found using (D.2) and (D.6)

$$\hat{\underline{\mu}} = \frac{1}{m} W \sum_{i=1}^{m} \underline{z}^i + \underline{\mu} \tag{D.11}$$

or

$$\hat{\underline{\mu}} = W\hat{\underline{\mu}}_{\underline{z}} + \underline{\mu} \tag{D.12}$$

For $\hat{\Sigma}$ can be written after substitution of (D.6) and (D.12) in (D.3)

$$\hat{\Sigma} = \frac{1}{m} \sum_{i=1}^{m} (W\underline{z}^i + \underline{\mu})(W\underline{z}^i + \underline{\mu})^T - (W\hat{\underline{\mu}}_{\underline{z}} + \underline{\mu})(W\hat{\underline{\mu}}_{\underline{z}} + \underline{\mu})^T \tag{D.13}$$

After some calculations it appears that (D.13) is equivalent to

$$\hat{\Sigma} = W\hat{\Sigma}_{\underline{z}} W^T \tag{D.14}$$

with

$$\hat{\Sigma}_{\underline{z}} = \frac{1}{m} \sum_{i=1}^{m} \underline{z}^i \underline{z}^{iT} - \hat{\underline{\mu}}_{\underline{z}} \hat{\underline{\mu}}_{\underline{z}}^T \tag{D.15}$$

For (D.14) it follows that

$$|\hat{\Sigma}|^{\frac{1}{2}} = \{|W||\hat{\Sigma}_{\underline{z}}||W|\}^{\frac{1}{2}} = |W||\hat{\Sigma}_{\underline{z}}|^{\frac{1}{2}} \tag{D.16}$$

Substitution of (D.5) - (D.10),(D.12), (D.14) and (D.16) in (D.4) yields after some calculations

$$E_{\chi}(e) = 1 - \int_{\underline{z}^1} \int_{\underline{z}^2} \cdots \int_{\underline{z}^m} \int_{\underline{z}} \min\{(2\pi)^{-k/2} |\hat{\Sigma}_{\underline{z}}|^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\underline{z}-\underline{\mu}_{\underline{z}})^T \Sigma_z^{-1} (\underline{z}-\hat{\underline{\mu}}_z)),$$

$$(2\pi)^{-k/2} \exp(-\tfrac{1}{2}\underline{z}^T\underline{z})\} \prod_{i=1}^{m} \{(2\pi)^{-k/2} \exp(-\tfrac{1}{2}\underline{z}^{i\,T}\underline{z}^i)\}$$

$$d\underline{z}\ d\underline{z}^1\ d\underline{z}^2 \cdots d\underline{z}^m \qquad (D.17)$$

This is independent of $\underline{\mu}$ and Σ.

APPENDIX E

THE MONTE CARLO PROCEDURE USED FOR THE ESTIMATION ERROR

Here the Monte Carlo method will be presented by which

$$e = 1 - \int_{\underline{x}} \min\{f_1(\underline{x}),\ f_2(\underline{x})\}\ d\underline{x} \tag{E.1}$$

is estimated. This is done by generating ν points $\underline{x}$ according to the density functions $f_1(\underline{x})$ and $f_2(\underline{x})$ and counting the number of times that $f_1(\underline{x})$ respectively $f_2(\underline{x})$ is smaller. The estimate $\hat{e}$ is given by

$$\hat{e} = 1 - \frac{1}{\nu}\sum_{i=1}^{\nu} \delta(\underline{x}_1^i) - \frac{1}{\nu}\sum_{i=1}^{\nu} (1 - \delta(\underline{x}_2^i)) \tag{E.2}$$

in which $\underline{x}_1^i$ $(i = 1,\nu)$ is randomly generated according to $f_1(\underline{x})$ and $\underline{x}_2^i$ $(i = 1,\nu)$ according to $f_2(\underline{x})$. The function $\delta(\underline{x})$ is defined as

$$\begin{aligned} \delta(\underline{x}) &= 1 \quad \text{if} \quad f_2(\underline{x}) > f_1(\underline{x}) \\ \delta(\underline{x}) &= 0 \quad \text{if} \quad f_2(\underline{x}) \leqslant f_1(\underline{x}) \end{aligned} \tag{E.3}$$

First will be proved that the estimator (E.2) is unbiased

$$E(\hat{e}) = 1 - E(\delta(\underline{x})|f_1) - 1 + E(\delta(\underline{x})|f_2) \tag{E.4}$$

$$= - \text{Prob}(\delta(\underline{x})=1|f_1) + \text{Prob}(\delta(\underline{x})=1|f_2) \tag{E.5}$$

$$= 1 - \{\text{Prob}(\delta(\underline{x})=1|f_1) + \text{Prob}(\delta(\underline{x})=0|f_2)\} \tag{E.6}$$

$$= 1 - \int_{\underline{x}} \{\delta(\underline{x})f_1(\underline{x}) + (1-\delta(\underline{x}))f_2(\underline{x})\}d\underline{x} \tag{E.7}$$

$$= 1 - \int_{\underline{x}} \min\{f_1(\underline{x}), f_2(\underline{x})\}\, d\underline{x} \tag{E.8}$$

$$= e \tag{E.9}$$

The variance of $\hat{e}$ follows easily

$$\mathrm{Var}(\hat{e}) = \frac{1}{\nu}\{\eta_1(1-\eta_1) + \eta_2(1-\eta_2)\} \tag{E.10}$$

in which

$$\eta_1 = \mathrm{Prob}(\delta(\underline{x})=1|f_1) \tag{E.11}$$

and

$$\eta_2 = \mathrm{Prob}(\delta(\underline{x})=1|f_2) \tag{E.12}$$

From (E.5) and (E.9) it follows that

$$\eta_2 - \eta_1 = e \tag{E.13}$$

or

$$\eta_2 = e + \eta_1 \tag{E.14}$$

Substitution in (E.10) yields

$$\mathrm{Var}(\hat{e}) = \frac{1}{\nu}\{\eta_1(1-\eta_1) + (e+\eta_1)(1-e-\eta_1)\} \tag{E.15}$$

$$= \frac{1}{\nu}\{\eta_1 - \eta_1^2 + e - e^2 - e\eta_1 + \eta_1 - \eta_1 e - \eta_1^2\} \tag{E.16}$$

$$= \frac{1}{\nu}\{2\eta_1 - 2\eta_1^2 - 2e\eta_1 + e - e^2\} \tag{E.17}$$

This is maximum for $\eta_1 = \frac{1}{2}(1-e)$, so

$$\mathrm{Var}(\hat{e}) \leqslant \frac{1}{2\nu}(1 - e^2) \tag{E.18}$$

is an upperbound for the variance of $\hat{e}$.

LIST OF MAIN SYMBOLS

A	class of objects
B	class of objects
c	a priori probability of class A
d	$\log \{(1-c)/c\}$
e	error in an estimate of a density function; estimation error
e_ℓ	estimation error of class ℓ
E	expectation operator
E_X	expectation operator over the learning set
$E_{\underline{\theta}}$	expectation operator over a family of density functions generated by a distribution over $\underline{\theta}$
$f(\cdot)$	density function
$f_\ell(\cdot)$	density function of class ℓ
$f_\ell^j(\cdot)$	density function of class ℓ for feature j
$g(\cdot)$	joint density of the learning objects
$g_\ell(\cdot)$	joint density of the learning objects of class ℓ
$g_{\underline{\theta}}(\cdot)$	a posteriori density of the parameters
$g_{\underline{\theta}_\ell}(\cdot)$	a posteriori density of the parameters of class ℓ
h	smoothing parameter in a Parzan estimation
$h(\cdot)$	a priori density of the parameters
$h_\ell(\cdot)$	a priori density of the parameters of class ℓ
I	identity matrix
k	number of features; feature size
m	number of learning objects of one class; sample size
n	number of cells; measurement complexity
p	probability that x takes on the value one

p_ℓ	probability that x takes on the value one, if $x \in$ class ℓ
p_ℓ^j	probability that x_j takes on the value one if $\underline{x} \in$ class ℓ
r^j	contribution of a single feature to $R(\underline{x})$ in case of independent features
$R(\cdot)$	discriminant function
$S(\cdot)$	discriminant function
$u(\cdot)$	kernel function
$v_\ell^j(\cdot)$	joint density of x_j and χ if $\underline{x} \in$ class ℓ
x	arbitrary one-dimensional object
$\underline{x}$	arbitrary k-dimensional object
x_j	feature value of feature number j of $\underline{x}$
$\underline{x}_\ell^i$	learning object number i of class ℓ
ε	classification error
ε^*	minimum value of ε; Bayes error
$\overline{\varepsilon}$	$E_\chi(\varepsilon)$; expected classification error
$\tilde{\varepsilon}$	$E_{\underline{\theta}}(\overline{\varepsilon})$; mean classification error
$\underline{\mu}$	expectation of a multivariate density
Σ	covariance matrix
$\underline{\theta}$	parameter vector $\underline{\theta} = \underline{\theta}_A : \underline{\theta}_B$
$\underline{\theta}_\ell$	parameter vector associated with the density of class ℓ
$\underline{\theta}_\ell^j$	parameter vector associated with the density of class ℓ for feature j.
χ	union of χ_A and χ_B
χ_ℓ	learning set of class ℓ
$\hat{a}$	estimation of a

REFERENCES

1. K. ABEND, T.J. HARLEY Jr., B. CHANDRASEKARAN, G.F. HUGHES, Comments "On the mean accuracy of statistical pattern recognizers", *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 420-423, 1969.

2. J. AITCHINSON, J.D.F. HABBEMA, J.W. KAY, A critical comparison of two methods of statistical discrimination, *Applied Statistics*, vol. 26, pp. 15-25, 1977.

3. D.C. ALLAIS, The problem of too many measurements in pattern recognition and prediction, *IEEE Int. Con. Rec.*, Part 7, pp. 124-130, 1966.

4. T. L. BOULLION, P.L. ODELL, B.S. DURAN, Estimating the probability of misclassification and variate selection, *Pattern Recognition*, vol. 7, pp. 139-145, 1975.

5. B. CHANDRASEKARAN, Independence of measurements and the mean recognition accuracy, *IEEE Trans. Comput.*, vol.IT-17, pp. 452-456, 1971.

6. B. CHANDRASEKARAN, A.K. JAIN, Quantization complexity and independent measurements, *IEEE Trans. Comput.*, vol. C-23, pp. 102-106, 1974.

7. B. CHANDRASEKARAN, A.K. JAIN, Independence, measurement complexity and classification performance, *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-5, pp. 240-244, 1975.

8. B. CHANDRASEKARAN, A.K. JAIN, "Independence, measurement complexity and classification performance": an emendation, *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-7, pp. 564-566, 1977.

9. T.M. COVER, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, *IEEE Trans. Elec. Comp.*, vol. EC-14, pp. 326-334, 1965.

10. P.A. DEVIJVER, Relationships between statistical risks and the least-mean-square-error design criterion in pattern recognition, *Proc. of the 1st Int. Joint Conf. on Pattern Recognition*, Washington, D.C., 1973.

11. R.O. DUDA, P.E. HART, Pattern classification and scene analysis, Wiley, New York, 1973.

12. R.P.W. DUIN, A criterion for the smoothing parameter for Parzen estimators of probability density functions, Internal report, Dept. of Applied Physics, Delft Univ. of Technology, 1975.

13. R.P.W. DUIN, On the choice of smoothing parameters for Parzen estimators of probability density functions, *IEEE Trans. Comput.*, vol. C-25, pp. 1175-1179, 1976.

14. R.P.W. DUIN, A sample size dependent error bound, *Proc. of the 3rd Int. Joint Conf. on Pattern Recognition*, Coronado, 1976.

15. R.P.W. DUIN, Comment on "Independence, measurement complexity and classification performance", *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-7, pp. 559-560, 1977.

16. R.P.W. DUIN, The mean recognition performance for independent distributions, *IEEE Trans. Inform. Theory*, In print.

17. O.J. DUNN, Some expected values for probabilities of correct classification in discriminant analysis, *Technometrics*, vol. 13, pp. 345-353, 1971.

18. B. EFRON, Biased versus unbiased estimation, *Advances in Mathematics*, vol. 16, pp. 259-277, 1975.

19. B. EFRON, C. MORRIS, Stein's estimation rule and its competitors - an emperical Bayes approach, *Journal of the Am. Statist. Ass.*, vo. 68, pp. 117-130, 1973.

20. B. EFRON, C. MORRIS, Data analysis using Stein's estimator and its generalizations, *Journal of the Am. Statist. Ass.*, vol. 70, pp. 311-319, 1975.

21. B. EFRON, C.MORRIS, Stein's paradox in statistics, *Scientific American*, vol. 26, no. 5, pp. 119-127, 1977.

22. H. FOLEY, Considerations of sample and feature size, *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 618-626, 1972.

23. K. FUKUNAGA, Introduction to statistical pattern recognition, Academic Press, New York, 1972.

24. G.F. HUGHES, On the mean accuracy of statistical pattern recognizers, *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 55-63, 1968.

25. L. KANAL, B. CHANDRASEKARAN, On dimensionality and sample size in statistical pattern classification, *Pattern Recognition*, vol. 3, pp. 225-234, 1971.

26. S. KULLBACK, Information theory and statistics, Wiley, New York, 1959.

27. G. J. LACHLAN, The asymptotic distributions of the conditional error rate and risk in discriminant analysis, *Biometrika*, vol. 61, pp. 131-135, 1974.

28. T. LISSACK, K.S. FU, Error estimation in pattern recognition via L^{α}-distance between posterior density functions, *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 34-45, 1976.

29. W. MOLENAAR, Approximations to the Poisson, bionomial and hypergeometric distributions, Mathematical Centre tracts 31, Amsterdam, 1973.

30. B.J.B. MOSS, An application of information theory in geological mapping, Thesis, Imperical College, Univ. of London, 1974.

31. M. OKAMOTO, An asymptotic expansion for the distribution of the linear discriminant function, *Ann. Math. Statist.*, vol. 34, pp. 1286-1301, 1963. Correction in *Ann. Math. Statist.*, vol. 39, p. 1358, 1968.

32. E.A. PATRICK, Fundamentals of pattern recognition, Prentice-Hall, Englewood Cliffs, 1972.

33. K.R. POPPER, Conjectures and refutations, Harper and Row, New York, 1968.

34. K.R. POPPER, Objective knowledge, Oxford University Press, Oxford, 1972.

35. S. RAUDYS, On dimensionality, learning sample size and complexity of classification algorithms, *Proc. of the 3rd Int. Joint Conf. on Pattern Recognition*, Coronado, 1976.

36. S. RAUDYS, Limitations of sample size in classification problems (in Russian), Institute of Physics and Mathematics of the Academy of Sciences of the Lithuanian SSR, Vilnius, 1976.

37. W.M.M. SCHINKEL, On the use of dependent binary features in the Bayes discriminant method (in Dutch), Thesis, Dept. of Applied Physics, Delft Univ. of Technology, 1975.

38. M. SORUM, Estimating the expected and the optimal probabilities of misclassification, *Technometrics*, vol. 14, pp. 935-943, 1972.

39. D. SPECHT, Generation of polynomial discriminant functions for pattern recognition, Technical report no. 6764-5, Stanford University, 1966.

40. R. STEINER, Die Philosophie der Freiheit, Rudolf Steiner Verlag, Dornach.

41. G.T. TOUSSAINT, Bibliography on estimation of misclassification, *IEEE Trans. Inform. Theory*, vol. IT-20, no. 4, pp. 472-47, 1974.

42. A.M. TURING, Computer machinery and intelligence, *Mind*, vol. 59, pp. 433-460, 1950.

43. J.R. ULLMANN, Experiments with the n-tuple method of pattern recognition, *IEEE Trnas. Comput.*, vol. C-18, pp. 1135-1137, 1969.

44. J. VAN NESS, C. SIMPSON, On the effects of dimension in discriminant analysis, *Technometrics*, vol. 18, pp. 175-187, 1976.

45. J.W. VAN NESS, Dimensionality and classification performance with independent coordinates, *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-7, pp. 560-564, 1977.

46. G.N. VAN VARK, A critical evaluation of the application of multivariate statistical methods to the study of human populations from their skeletal remains, *Homo*, Band 27, Heft 2, pp. 94-114, 1977.

ACKNOWLEDGEMENT

It is a pleasure for me to thank all the co-workers and students of the Pattern Recognition Group of the Department of Applied Physics of the Delft University of Technology for creating the stimulating environment necessary for my work on statistical pattern recognition.

I am obliged to mrs. B.J.M. Scholten - van der Burg and mrs. S. Kranenburg - Ginjaar, who typed the manuscript, to mr. A.S.G. de Knegt, who drew the diagrams, to mr. S. Lobregt, who assisted in preparing the front page, and to miss J. Warris and mr. L.J. Kraamer, who translated parts of the Russian book by Raudys.

STELLINGEN

behorende bij het proefschrift van

R.P.W. Duin

Delft, 14 juni 1978

1 De belangrijkste voorwaarde voor de ontwikkeling van een nauwkeurig patroonherkennend systeem is de selectie op niet-statistische gronden van een gering aantal bruikbare kenmerken. De beschikbaarheid van een grote verzameling leerobjecten verzacht deze eis enigszins, doch niet essentieel.

2 Bij het ontwikkelen van een patroonherkennend systeem zal voortdurende vergroting van het aantal in te stellen parameters het herkenningsresultaat uiteindelijk doen dalen. Uitsluitend op grond van een verzameling leerobjecten valt niet te bepalen wanneer dit 'piekeffect' optreedt.

3 Het model van Hughes is vanwege zijn te universele opzet ongeschikt voor een algemene studie van het 'piekeffect'.

Hughes, IEEE Trans. Inf. Theory, IT-14, 55-63, (1968).

4 Aan het in dit proefschrift behandelde 'piekeffect' is een argument te ontlenen tegen de bewering van Turing dat een computer, mits juist geprogrammeerd, niet te onderscheiden is van een mens op grond van zijn waarneembaar gedrag.

Turing, Mind, 59, 433-460, (1950).

5 Het meermalen gebruiken van dezelfde verzameling testobjecten bij de ontwikkeling van een patroonherkennend systeem veroorzaakt te optimistische foutschattingen.

6 De moeilijkheid een synthese te bewerkstelligen tussen het linguistisch en het statistisch patroonherkennen is inherent aan de complementariteit van de deterministische en de probabilistische benaderingswijzen in de natuurwetenschappen.

7 De mogelijkheid tot het volgen van een college over de door Goethe gebruikte methode voor natuurwetenschappelijk onderzoek behoort in een academische opleiding in de natuurkunde niet te ontbreken.

8 Bestudering van levensprocessen uitsluitend met behulp van instrumenten ontwikkeld bij de studie van de levenloze natuur leidt tot een eenzijdige, materialistische visie op deze processen.

9 De ontwikkeling van het probabilistische denken in de natuurkunde vertoont historisch gezien een nauwe samenhang met de manier waarop het deterministische denken zich daar heeft ontwikkeld.

Hacking, The emergence of probability, Cambridge University Press (1975).

10 Er zijn duidelijke argumenten aan te voeren, o.a. van embryologische aard, tegen de evolutieleer van Darwin en zijn moderne varianten.

Poppelbaum, Mensch und Tier, Philosophische Verlag Dornach.

11 Bij het onderwijs in de statistiek dient het onderscheid tussen het bij axioma ingevoerde toeval en de werkelijkheid te worden benadrukt.

12 Automatische herkenning van nummerborden langs de rijkswegen is zonder duidelijke wettelijke waarborgen voor de bescherming van de privé-sfeer van de burger ongewenst.