ELECTRICAL SCIENCE SERIES

INTRODUCTION TO STATISTICAL PATTERN RECOGNITION

ELECTRICAL SCIENCE
A Series of Monographs and Texts

Editors: Henry G. Booker and Nicholas DeClaris
A complete list of titles in this series appears at the end of this volume.

INTRODUCTION TO STATISTICAL PATTERN RECOGNITION

Keinosuke Fukunaga

SCHOOL OF ELECTRICAL ENGINEERING
PURDUE UNIVERSITY
LAFAYETTE, INDIANA

ACADEMIC PRESS New York and London 1972

ACADEMIC PRESS, INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1 7DD

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 72-75627

PRINTED IN THE UNITED STATES OF AMERICA

To Reiko, Gen, and Nina

This page intentionally left blank

CONTENTS

Chapter 8 Feature Selection and Linear Mapping for One Distribution

Chapter 9 Feature Selection and Linear Mapping for Multidistributions

Chapter 10 Nonlinear Mapping

Chapter 11 Clustering

This page intentionally left blank

PREFACE

This book presents an introduction to statistical pattern recognition. Pattern recognition in general covers a wide range of problems, and it is hard to find a unified view or approach. It is applied to engineering problems, such as character readers and waveform analysis, as well as to brain modeling in biology and psychology. However, statistical decision and estimation, which are the subjects of this book, are regarded as fundamental to the study of pattern recognition. Statistical decision and estimation is covered in various texts on mathematical statistics, statistical communication, control theory, and so on. But obviously each field has a different need and view. So that workers in pattern recognition need not look from one book to another, this book is organized to provide the basics of these statistical concepts from the viewpoint of pattern recognition.

The material of this book has been taught in a first-level graduate course at Purdue University and also in a special summer course at IBM, Rochester, Minnesota. Therefore, it is the author's hope that this book will serve as a text for the introductory courses of pattern recognition as well as a reference book for the workers in the field.

One difficulty in pattern recognition is that we have to handle a large number of correlated random variables. This leads us to rely heavily on linear algebra. In Chapter 2, a survey of linear algebra is included with a review of the properties of random variables and vectors. Throughout the book,

particular emphasis is placed on viewing the problems in terms of the eigenvalues and eigenvectors.

In Chapters 3–7, classifier design is discussed. In addition to the standard material of hypothesis testing (Chapter 3) and parameter estimation (Chapter 5), the estimation of the error probability is emphasized in these chapters. The probability of error is the key parameter in pattern recognition. Chapter 4 is devoted to linear and piecewise linear classifiers because they often are the only classifiers that can be practically implemented. Another difficulty in pattern recognition is that the normal (Gaussian) assumption does not hold for most applications. Because of this fact, a nonparametric approach to the problem is unavoidable in practice (Chapter 6). Chapter 7 discusses successive approaches where the classifier is adaptively adjusted each time one sample is observed.

In Chapters 8–10, feature selection is discussed from the viewpoint of mapping the original measurement space into a lower-dimensional feature space, without losing the information of our interest. Linear mappings are applied to select a set of features which minimizes the error of representing samples from one distribution (Chapter 8) or maximizes the class separability for multidistributions (Chapter 9). The discussion is then extended to include nonlinear mappings (Chapter 10).

Chapter 11 is devoted to clustering or unsupervised classification where samples are classified with a minimum of *a priori* knowledge about their distribution.

ACKNOWLEDGMENTS

The author would like to express his gratitude to Dr. J. C. Hancock and his colleagues at Purdue University for their encouragement. Also, it is the author's pleasure to acknowledge the support of the National Science Foundation for research in pattern recognition. Much of the material in this book was contributed by the author's past and present co-workers, Mr. D. L. Kessell, Dr. W. L. G. Koontz, Dr. T. F. Krile, and Dr. D. R. Olsen. The author is particularly grateful to Dr. Koontz for his thoughtful and detailed criticism of the entire manuscript as well as his significant contribution to the content. In addition, the author wishes to thank his wife Reiko for her typing of the manuscript.

The author acknowledges the Institute of Electrical and Electronics Engineers, Inc., The Institute of Mathematical Statistics, and the American Telephone and Telegraph Co. for their authorization to use material from their journals.

This page intentionally left blank

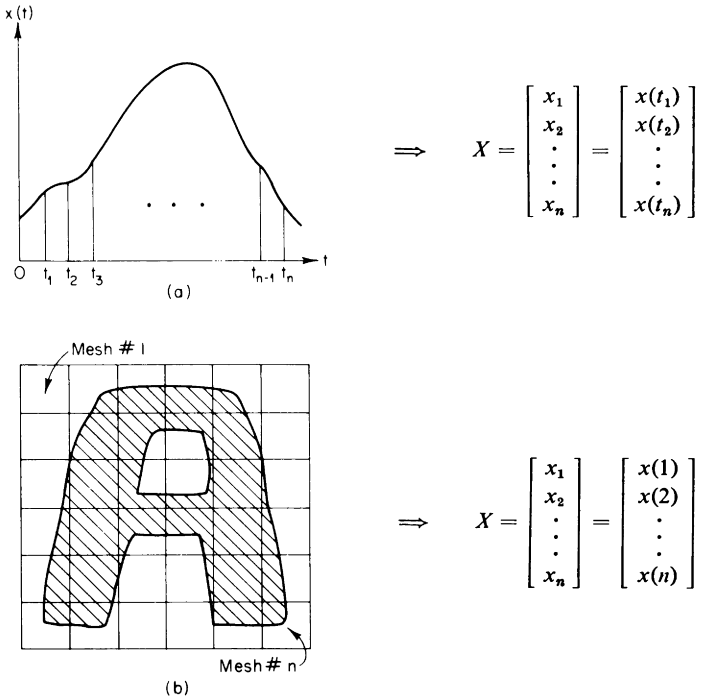
Chapter 1

INTRODUCTION

This book presents and discusses the fundamental mathematical tools for statistical decision-making processes in pattern recognition. It is felt that the decision-making processes of a human being are somewhat related to the recognition of patterns; for example, the next move in a chess game is made based on the present pattern on the board, and the buying or selling of stocks is decided by a complex pattern of information. Therefore, the goal of pattern recognition is to clarify these complicated mechanisms of decision-making processes and to automate these functions using computers. However, because of the complex nature of the problem, most pattern recognition research has been concentrated on more realistic problems, such as the recognition of Latin characters and the classification of waveforms. The purpose of this book is to cover the mathematical models of these practical problems and to provide the fundamental mathematical tools necessary for solving them. Although many approches have been proposed to formulate more complex decision-making processes, these are outside of the scope of this book.

1.1 Formulation of Pattern Recognition Problems

Many important applications of pattern recognition can be characterized as either waveform classification or classification of geometric figures. For example, consider the problem of testing a machine for normal or abnormal operation by observing the output voltage of a microphone over a period of time. This problem reduces to discrimination of waveforms from good and bad machines. On the other hand, recognition of printed English characters corresponds to classification of geometric figures. In order to perform this type of classification, we must first measure some observable characteristics of the sample. The most primitive way to do this is to measure the time-samples values for a waveform, $x(t_1),\ldots,x(t_n)$, and the grey levels of meshes for a figure, $x(1),\ldots,x(n)$, as shown in Fig. 1-1. These n measurements form a vector, X. Even under the normal machine condition, the observed waveforms are different each time the observation is made. Therefore, $x(t_i)$ is a random variable and will be expressed, using



$$\Rightarrow \quad X = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} x(t_1) \\ x(t_2) \\ \cdot \\ \cdot \\ \cdot \\ x(t_n) \end{bmatrix}$$

$$\Rightarrow \quad X = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} x(1) \\ x(2) \\ \cdot \\ \cdot \\ \cdot \\ x(n) \end{bmatrix}$$

Fig. 1-1 Two measurements of patterns: (a) waveform; (b) character.

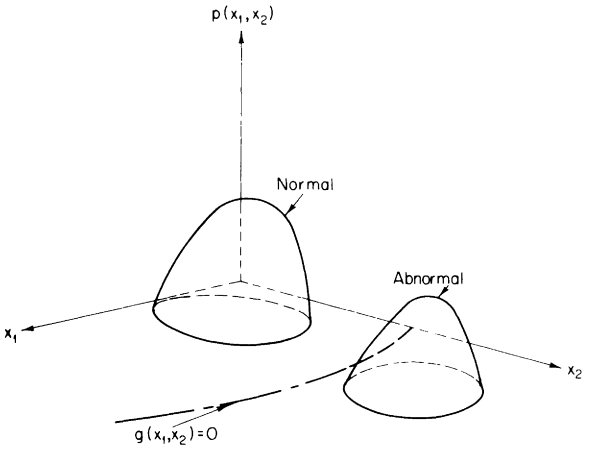


Fig. 1-2 Distributions of **X** for normal and abnormal conditions.

boldface, as $\mathbf{x}(t_i)$. Likewise, X is called a random vector if its components are random variables and is expressed as **X**. Similar arguments hold for characters: the observation, $x(i)$, varies from one A to another and therefore $\mathbf{x}(i)$ is a random variable, and **X** is a random vector.

Thus each waveform or character is expressed by a vector in an n-dimensional space, and many waveforms or characters form a distribution of **X** in the n-dimensional space. Figure 1-2 shows a simple two-dimensional example of two distributions corresponding to normal and abnormal machine conditions. If we know these two distributions of **X** from past experience, we can set up a boundary between these two distributions, $g(x_1, x_2) = 0$, which divides the two-dimensional space into two regions. Thus, when an unknown waveform is observed, we can decide whether the waveform comes from a normal or abnormal machine, depending on $g(x_1, x_2) < 0$ or $g(x_1, x_2) > 0$. We call $g(x_1, x_2)$ a discriminant function, and a network which detects the sign of $g(x_1, x_2)$ is called a pattern recognition network, a categorizer, or a classifier. Figure 1-3 shows a block diagram of a classifier in a general n-dimensional space. Thus, in order to design a classifier, we must study the characteristics of the distribution of **X** for each category and find a proper discriminant function.

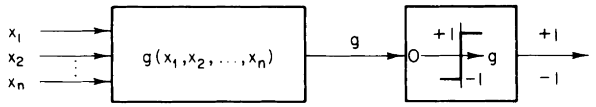


Fig. 1-3 Block diagram of a classifier.

In the above discussion we assumed a very primitive way of choosing measurements. Since each of these primitive measurements carries a very small amount of information about the sample, the number of measurements n usually becomes high, perhaps in the hundreds. This high dimensionality makes many pattern recognition problems difficult. On the other hand, classification by a human being is usually based on a small number of features, such as the peak value, fundamental frequency, etc. Each of these measurements carries significant information for classification purposes and is selected according to the physical meaning of the problem. Obviously, as the number of inputs to a classifier becomes smaller, the design of the classifier becomes simpler. In order to enjoy this advantage, we have to find some way to select or extract important features from the observed samples. This problem is called feature selection or extraction and is another important subject of pattern recognition. Feature selection can be considered as a mapping from the primitive n-dimensional space to a lower dimensional space. The class separability of the distributions of the primitive measurements is the same as that of the samples. Therefore, the mapping should be carried out without severely reducing this class separability.

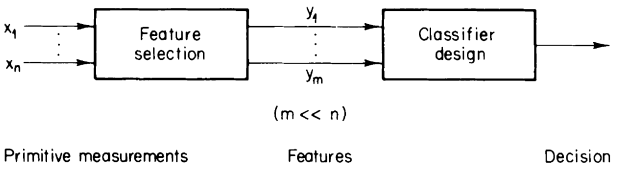


Fig. 1-4 Block diagram of pattern recognition.

Thus, as shown in Fig. 1-4, pattern recognition consists of two parts: feature selection and classifier design. In practice, there are no clear ways to separate these two operations. In fact, the classifier can be viewed as a feature selector which maps m features to one feature (the discriminant function). However, it is convenient to separate the problem into two parts and to study them independently.

1.2 Chapter Outlines

In order that pattern recognition be studied within the scope mentioned above, this book has been divided into ten chapters (2–11).

In *Chapter 2*, the various properties of random vectors and the methodology of linear algebra are surveyed. The knowledge of this material is

required for a complete understanding of this book. However, it is assumed that the reader is familiar with the properties of a random variable and a random vector, so only a quick survey is presented here. Also, since vectors and matrices are used extensively throughout the book, linear algebra is surveyed, particularly from the viewpoint of eigenvalues and eigenvectors.

Chapter 3 through 7 are related to classifier design.

In *Chapter 3*, we seek the best theoretical way to design a classifier, assuming that the distributions of the random vectors to be classified are given. In this case, the problem becomes simple statistical hypothesis testing. The Bayes classifier is derived as optimum in that it minimizes the probability of error of classification or the risk under preassigned costs for various decisions. The Neyman–Pearson test and the minimax test are also introduced.

The probability of error is the key parameter in pattern recognition. It is the measure of the class separability of given distributions, if we assume the use of the Bayes classifier. Also, it is the measure of the performance of a classifier in comparison with the Bayes classifier for given distributions. Because of its importance, in Chapter 3, we discuss how to calculate the probability of error for given distributions. We also consider the simpler problem of finding an upper bound of the error probability.

In an alternate formulation of the pattern recognition problem, a sequence of samples from the same class is observed. It is well known that the class can be determined with considerably greater assurance by observing a sequence of samples rather than a single sample. Therefore, the sequential hypothesis test is also included.

In *Chapter 4*, linear classifiers are explored. Although the Bayes classifier is optimal, its implementation is often difficult in practice because of its complexity, particularly when the dimension is high. Therefore, we are often led to consider a simpler classifier. Linear or piecewise linear classifiers are the simplest and most common choices. Various design procedures for linear classifiers are discussed in Chapter 4. These include the Bayes classifier for certain distributions, the optimum linear classifier in the sense of the minimum probability of error or in the sense of the minimum mean-squared error, and so on. The case where the input measurements are binary is also considered.

In *Chapter 5*, parameter estimation is discussed. In the earlier chapters, we assume that the distributions to be classified are given. However, in practice, only a finite number of samples is available, and we have to estimate the distributions from these samples. When the functional form of a distribution is known, the density function can be estimated by replacing unknown