Book Reviews

Pattern Classification and Scene Analysis—Richard O. Duda and Peter E. Hart (New York: Wiley-Interscience, 1973, 482 pp., $22.50).

Introduction to Statistical Pattern Recognition—Keinosuke Fukunaga (New York: Academic Press, 1972, 369 pp., $18.50).

Fundamentals of Pattern Recognition—Edward A. Patrick (Englewood Cliffs, N.J.: Prentice-Hall, 1972, 504 pp., $18.00).

Introduction to Mathematical Techniques in Pattern Recognition—Harry C. Andrews (New York: Wiley-Interscience, 1972, 242 pp., $11.50).

Computer-Oriented Approaches to Pattern Recognition—William S. Meisel (New York: Academic Press, 1972, 250 pp., $15.00).

THOMAS M. COVER

$85.50 is the price for the 1,847 pages of reading in pattern recognition contained in the books under review. Except for collections of papers and a book of solicited chapters edited by Mendel and Fu, these are the first texts in statistical and geometric pattern recognition since Nilsson's book *Learning Machines*, published in 1965. So it is with some interest and concern that we examine this multiple offering.

Work on pattern recognition properly got its start in the late 1950's, primarily due to the work of Rosenblatt and Block on the Perceptron at Cornell University and Selfridge and colleagues at M.I.T. There followed a book by Sebestyen and an influential paper by Highleyman. The first book on the general subject of pattern recognition appeared in 1965. This was Nilsson's book, *Learning Machines*. The primary emphasis in this book was on discriminant functions and the Perceptron, Adaline, linear threshold device approach. This simple but comprehensive book made clear some of the fundamental ideas that had arisen in pattern recognition in the preceding years. While it cannot be said that any of these techniques allowed the solution of the pattern recognition problem, it was clear that the general problem of recognizing patterns had inspired new approaches to some old mathematical problems in hypothesis testing and statistics. In 1969, the book *Perceptrons* by Minsky and Papert appeared. The book is by far the deepest and most entertaining entry in the field of pattern recognition to date. However, it deals specifically with linear threshold devices and what the authors call the emerging field of computational geometry. This book should be part of the cultural background of any serious researcher in pattern recognition, but I seriously doubt whether it will be of any direct utility. What Minsky and Papert succeeded in emphasizing was that given a device, say the Perceptron, one should find the natural family of problems for which it is suited. Somehow in the previous years, it had been assumed that ad hoc devices were somehow universally good, and no attempt was made to delineate the natural family of problems for which they were suited. Thus the change in philosophy was extremely important.

In any case, except for the book by Minsky and Papert, the book by Nilsson, an excellent text by Rosenfeld on pictorial pattern recognition, there has been little except for volumes of collected papers published in pattern recognition. Now we find that five books have simultaneously come upon the scene. These books will be reviewed here.

Preview

It is fairly clear what the words "pattern" and "recognition" mean and therefore what the central aim of pattern recognition is—namely, a systematic theoretical and experimental attempt to develop simple computational means for placing abstract objects into categories, with the eye-brain computer as one (as yet unsurpassed) model.

The books under review have personalities. The overwhelming impression gathered from Duda and Hart's book is that it is a work of careful scholarship. Chapters end with a dense raisin-cake of references, dates, and cross-references. The history of the subject is gathered together in a manner that would do justice to a good current review article on the subject. I am reminded in particular of the review article of Nagy in 1968, "State of the Art in Pattern Recognition."

An illusory feeling of well-being comes from reading Andrews' book. It is literate and interestingly written and is also the beneficiary of a very good layout, but the treatment is not deep. Andrew's book is more like an outline for a short course than a serious mathematical development of the subject.

The book by Meisel has much in common with that of Andrews, both in subject matter and style. Even some of the problems are the same. This may be because Meisel originated a course at U.S.C. which Andrews subsequently became involved in. Although Andrews had the second crack at the subject, Meisel's book seems to be the more ambitious and successful effort.

The exposition in Patrick's book is almost unreadable, and the notation is awful. But Patrick's book is an honest effort, although careless and somewhat lacking in discrimination. Patrick is locked in a losing struggle with notation and clarity. But, while Fukunaga and Duda and Hart have succeeded in this struggle to some extent, Andrews and Meisel have avoided it almost entirely. Thus Patrick earns my admiration because, above all, he loves the subject, and his effort shows.

Fukunaga has a nice feeling for the subject, but is very heavy on classical multivariate techniques and tends not to view pattern recognition as a unique area, but primarily as a domain of classical statistics. The book is good in pattern recognition but flawed according to its own standards by the shaky statistical exposition.

Pattern Classification and Scene Analysis—Richard O. Duda and Peter E. Hart.

The 482-page text by Duda and Hart is a dense and scholarly treatment of major advances in pattern recognition that have taken place over the last ten years. In addition, it gives a very thorough summary of the work presented in Nilsson's 1965 book, *Learning Machines*, as well as some of the new material appearing in Rosenfeld's book on pictorial pattern recognition. While the other books are primarily statistical in nature, this book splits nicely into two parts: first, a part primarily written by Duda that treats statistical pattern recognition; and the second part, primarily written by Hart, that treats computer and pictorial pattern recognition techniques. The subjects include Bayes decision theory, parameter estimation and supervised learning, nonparametric techniques, density estimation, linear discriminant functions, unsupervised learning and clustering, scene analysis, frequency domain approach, descriptions of line and shape, perspective transformations, projective invariants, and descriptive methods in scene analysis.

Duda and Hart give a very thorough list of references and bibliography and, in addition, draw the reader's attention to many fine survey articles in the field. The Bayes decision theory is well known but is thoroughly and cleanly developed by the authors. In addition, the authors prove that independent binary features in the two category problem result in a Bayes decision surface which is a linear hyperplane —one of the basic motivating results for linear threshold devices when they were in vogue several years ago. This simple result is also given a good treatment in Fukunaga and in Minsky and Papert.

Dimensionality

Duda and Hart's discussion of problems of dimensionality is excellent and contains many refreshing ideas. The authors first discuss fitting polynomials to points and the tradeoff between fitting an mth degree polynomial to n points ($m > n$) with subsequent good fit but poor generalization, versus poorly fitting a lower degree polynomial to the n points, hopefully with subsequent better prediction of function

values on new points. They go on to find the capacity of a hyperplane; namely, two patterns per degree of freedom of the hyperplane. They then go on to give a brilliantly clear description of G. F. Hughes' result that the average probability of error when one divides an arbitrary feature space into m cells and has n samples, decreases to a minimum as m increases and then ascends to one half as the feature partitioning m goes to infinity. This is directly related to the difficulties inherent in classification when the feature set is too large for a fixed sample size. Although Hughes' discussion is clear, I believe the presentation in Duda and Hart is more succinct and has more impact.

Finally, in a short section on estimating the error rate, the authors restrict themselves to a discussion of estimating the error rate of a given rule on a separate test set—a much simpler problem than estimating the performance of a decision rule without the aid of additional test samples. Here they bring in confidence intervals (Highleyman, 1962), in what is a very useful presentation that should aid practicing statistical pattern recognizers in establishing confidence intervals for the error rates of their procedures. The references for this section do a fine job in attributing credit and characterizing the efforts of other workers in this area, including Chandrasekaran, Kanal, Lachenbruck, Fukunaga, Kessell, and Mickey.

Density Estimation

In density estimation, Duda and Hart discuss Loftsgaarden and Quesenberry's nearest-neighbor approach, and the Parzen–Rosenblatt estimators. Here we have a proof of the convergence of the Parzen estimate. The authors find and exhibit the correct conditions on the size of the window functions sufficient for convergence. The proof is somewhat heuristic but satisfying.

Some graphs illustrate the difference between the Parzen estimates and the Loftsgaarden–Quesenberry estimates; but, since they are on a log-density scale, the comparisons are of little value. The log density tends to infinity, and in the regions of low density for finite sample size the approximations will be very poor. As far as pattern-classification use is concerned, it is the behavior of the density estimates in the region of high probability density that is usually the most important.

However, the ultimate use of density estimation procedures in classification is not adequately exposed. The authors fail to mention the work of Van Ryzin and others showing that the Bayes decision rule with respect to consistent estimators (like the Parzen estimator) will yield a sequence of probabilities of error converging to the Bayes risk in the limit as n tends to infinity.

The proof that the nearest neighbor rule has a risk bounded above by twice the Bayes risk in the large-sample case again appeals to δ-function intuition. Although this is satisfying and certainly (with the exception of Fukunaga and Patrick) better than the nothing we find from the other authors, one cannot tell what the necessary conditions are for convergence. Nonetheless, Duda and Hart are extremely careful in delineating the nonrigorous sections of their presentation and no fair-minded reviewer can quibble with the patent simplicity of the presentation and its consequent pedagogical value.

Further novelties with respect to the existing pattern recognition literature are presented in the Rademacher–Walsh expansion, a full discussion of probability distributions on the binary n-cube, and the Bahadur–Lazarsfeld expansion. Finally the authors discuss the Chow expansion for probability distributions, in which first-order Markov-type tree dependence is assumed to exist among the variables.

To show the care of Duda and Hart, consider the bibliographical and historical remarks at the end of Chapter 4. The authors say:

> Our treatment of the Parzen window method is a slight generalization of the univariate formulation by Rosenblatt, 1956. Rosenblatt's work actually preceded that of Parzen, 1962, but Parzen had previously employed similar methods for the estimation of spectra, and the phrase, "Parzen window," is now well established.

This fair discussion of the past of the subject returns to Rosenblatt some of the needed credit for this nice density-estimation scheme. Throughout the book, the historical discussions are excellent and scholarly.

Linear Discriminant Functions

The chapter on linear discriminant functions is also excellent. Fully eight descent procedures for obtaining optimal linear discriminant functions are discussed, ranging from the fixed-increment (Perceptron) procedure to relaxation, stochastic approximation, Ho–Kashyap procedures, and linear programming. The amount of work involved in isolating the essence of each of these procedures is extraordinary. My only quibble is that the stochastic approximation procedure, while updating the parameter of the discriminant function, does not selectively choose arguments of the function to be evaluated and thereby does not require the full power of stochastic approximation.

The discussion of decomposition of mixtures of distributions and clustering is properly combined in Chapter 6 of Duda and Hart. The authors neglect measure-zero distinctions in defining identifiability. Patrick is more thorough here, and Fukunaga's section on clustering is equally comprehensive. Clustering is one of the fine contributions of pattern recognition, and I am not surprised to see that Duda and Hart, Patrick, and Fukunaga have all done an excellent job on it.

The discussion of unsupervised Bayesian learning is very nice in Duda and Hart. Finally, the discussion of graph-theoretic techniques and mean-squared-error criteria for clustering in the absence of any statistical information whatsoever is very good, and is replete with algorithms for which convergence to local minima can be established.

Incidentally, at about the time of the publication of these books, there appeared a book on clustering methods (*Mathematical Taxonomics* by Jardine and Sibson) which is entirely devoted to clustering but does not contain some of the material in these five books under review. However, an understanding of clustering would be incomplete without reference to Jardine and Sibson.

Pictorial Pattern Recognition

The second half of Duda and Hart enters an area which is not covered in the other four books under review; namely, that of pictorial pattern processing as opposed to statistical data processing. Here questions of spatial smoothing, edge detection, the determination of texture, template match, identification of picture components (is there a chair in the picture?), are considered. In some cases statistical analyses are given of the ad-hoc picture processing procedures, such as an analysis of template matching under a maximum-likelihood criterion. A discussion of normalized color coordinates is given.

At one point, the authors show "A Simple Scene" which consists of a photograph of a desk with a telephone on it and a picture behind it which is a picture of a desk with a telephone on it. It takes a few seconds to discover where to look to find out whether this is indeed an infinite regress.

There is a section on the spatial frequency domain that draws on the work of Papoulis, Bracewell, and an earlier book by Andrews, *Computer Techniques in Image Processing* (Academic Press, 1970). Here, sampling theorems and a discussion of template matching of the convolution theorem are given in a very readable discussion.

While none of the other books treat Minsky and Papert's beautiful results on the complexity of geometrical predicates as looked at by perceptrons, this book at least makes a preliminary presentation of these results—first showing the difficulty of various geometrical predicates for diameter-limited perceptrons and then going on to order-limited perceptrons. Most of the results are stated without proof.

There is a section on integral geometric descriptions. Most of the results are given without proof and the discussion of Bertrand's paradox is not illuminating.

The section on perspective transformations is very good and is also not presented in any of the other books. Much of this is needed background for any TV-monitored pattern recognition problems. And if stereo perception is needed, the material is essential. The cross ratio is introduced in the section on projective invariants. Finally some

descriptive formalisms and picture grammars are developed. Special applications to analyses of polyhedra are given and Huffman's nice analytical results on impossible objects are presented. Finally, there is a good presentation of Guzman's 1968 work on grouping picture regions into objects.

In summary, the book by Duda and Hart is an extremely scholarly and solid work, comprehensive in detail. It covers both statistical and pictorial techniques but does not skimp on effort in making the mathematical details lucid and precise.

Introduction to Statistical Pattern Recognition—Keinosuke Fukunaga.

At the time I taught my class this winter (1973), Duda and Hart's book was not yet out, nor was its applicability to the course yet clear in my mind, so Fukunaga became my selection for a reference for the course. We found the book to be stimulating and thorough.

Fukunaga concerns himself primarily with those aspects of pattern recognition that can be characterized as waveform classification or classification of geometric figures. This, of course, covers a lot of territory.

Probability Background

Chapter 2 of Fukunaga concerns random vectors and their properties and is basically a quick run-through of probability ideas at a level somewhat lower than Parzen's elementary text. There are some annoying errors. For example, the components of the covariance matrix are expressed as σ_{ij}^2 in three places on page 17 and once on 18, thereby more than strongly suggesting that all of the terms are nonnegative.

Also ambiguous is Fukunaga's notation on page 26, "If $P(X_N) \to P(Z)$ as $N \to \infty$, we say that $\mathbf{X}_N$ tends to $\mathbf{Z}$ in distribution." Here P is a distribution function, X_N, Z are real numbers, and X_N, $\mathbf{Z}$ are random variables. The definition makes little sense. On the same page the author states the central limit theorem as an example of convergence in distribution, but unfortunately normalizes $\sum_{i=1}^{n} X_i$ by $1/n$ instead of $1/\sqrt{n}$, thereby yielding a degenerate and distinctly nonnormal limit.

Fukunaga's discussion of the Neyman–Pearson test omits proof of the Neyman–Pearson lemma, does not discuss necessary randomization when the distributions have atoms, and, in fact, uses a Lagrange multiplier minimization approach which, strictly speaking, fails because the minimum is achieved at the boundary of the set of all decision rules. The needed Kuhn–Tucker theorem justification is simply the Neyman–Pearson lemma itself in this context. In general this treatment is unnecessarily murky.

Fukunaga gives the Ho–Kashyap algorithm and convergence proof for finding a separating hyperplane when the patterns are linearly separable. This is a gradient descent method. He also provides a proof, in a different chapter, of the convergence of the fixed increment algorithm. Finally, although Fukunaga's treatment of linear classifiers is not as complete as Nilsson, Duda and Hart, Andrews, and Meisel, he includes the important observation that many seemingly nonlinear discriminant functions can be considered to be linear in their coefficients and therefore linear discriminant functions in some space into which all the patterns have been mapped. This observation allows linear techniques to be used for polynomial discrimination problems.

Fukunaga's discussion of the sufficient statistic involves first a Bayesian characterization in terms of *a posteriori* distributions on the parameter, an unnecessarily restrictive point of view. Later, on page 130, he proves the factorization criterion for a sufficient statistic but does not discuss the concept of minimal sufficient statistic. Thus the discussion is needlessly incomplete and references to standard texts like Ferguson and Blackwell and Girshick are not provided.

Error Estimation

Fukunaga provides a nice treatment of estimation of the probability of error of a given decision rule based on independent, unlabelled test samples. He then goes on to discuss what happens when one designs the classifier on the training set and then uses the training set to estimate the probability of error of the classifier. This involves some sophisticated jack-knifing procedures; among them, the procedures of Hills and Lachenbruck. This area is of intense current interest. Fukunaga is one of the primary contributors in this area.

Density Estimation

In Chapter 6 Fukunaga, in the manner of Andrews and Meisel, considers estimation of density functions—but this is mostly a summary with proofs omitted. Unlike Andrews and Meisel, Fukunaga gets the details scrupulously correct and puts conditions on the kernel estimation procedures of Parzen and Rosenblatt which guarantee convergence in probability. In addition, Fukunaga proves Bochner's theorem, which is the underlying theorem for the establishment of consistency of Parzen-type estimators. He continues with material from Parzen's 1962 paper by proving uniform consistency of the estimation of the mode of an unknown density function—a very important procedure for pattern recognition. Conditions for convergence of the Loftsgaarden and Quesenberry procedure are provided.

Fukunaga's proof of the convergence of the risk of the nearest neighbor rule is given in outline form. Exchanges of limits and expectations are not justified (in this case by dominated convergence), while in the somewhat more rigorous, equally intuitive proof of Duda and Hart, this condition is mentioned. While I do not believe that simple theorems like Fubini's theorem need be invoked in a discussion at this level, dominated convergence is less often satisfied. For example, unawareness of this condition would have one assume that the same sorts of bounds hold for nearest-neighbor estimation procedures without additional conditions. This has proved not to be the case. Incidentally, only Patrick and, to some extent, Duda and Hart extend the classification procedures of the nearest neighbor rule to estimation procedures, even though everyone knows that classification is a special case of estimation; and many pattern recognition problems are really parameter estimation problems, as the authors amply imply by covering estimation material without exception in the books under review.

Stochastic Approximation

Fukunaga gives an interesting discussion of stochastic approximation at the outline level. In his development of the proof of the convergence of stochastic approximation, he states the weird implicit condition, p. 207, "We assume the regression function is also bounded in the region of our interest as $E\{f(\theta_N)^2\} \leq M$." If we are trying to prove some good properties of our sequence of estimates of roots θ_N, this constraint seems to me to be begging the question in an important way. On page 207 and again on 208, with respect to stochastic approximation, Fukunaga makes a nonsense statement about convergence with probability one, by implying that $\theta_i \to \theta$, wpl, if $\lim_{i\to\infty} \Pr\{\theta_i = \theta\} = 1$. This is not even convergence in probability. In fact, for θ_i continuous r.v.'s, $\Pr\{\theta_i = \theta\} \equiv 0, \forall_i$.

On page 222, Fukunaga makes an incisive comment about the distinction or lack thereof between unsupervised Bayesian learning and supervised Bayesian learning:

> Therefore, as a concept, successive unsupervised estimation ... is the same as successive supervised estimation. ... But because of the summation involved in the calculation of *a priori* density function, the reproducing property is lost for all density functions listed previously, including the normal distribution. This means that we cannot easily estimate a set of parameters so that we have to deal with the recursive estimation of multivariate functions.

In other words, as noted by Spragins, finite-dimensional sufficient statistics are not preserved in unsupervised estimation problems, because mixtures of distributions do not preserve the finite-dimensional sufficient statistics of the component distributions.

Feature Selection

In Chapter 8, Fukunaga discusses feature selection using linear mappings. This is essentially a statistical approach putting emphasis

on the Karhunen–Loève expansion. The primary idea here is to re-represent the data in a one-to-one way in some other space, hopefully a space in which the subsequent operations will be simple. This puts off the question of whether this preliminary data processing is indeed necessary.

Concerning feature selection, Fukunaga says:

> When one distribution is studied, there are no classification problems—only representation problems. It is assumed that features which represent individual distributions should lead us to good features for classifying these distributions. Feature selection for one distribution is a mapping from the original n-dimensional space to an m-dimensional space ($m \ll n$) which does not distort the representation of a given distribution. If the classification problem is not explicitly considered, we have no way to determine what kind of properties of a given distribution should be preserved.

This perhaps explains why so many authors have concerned themselves with the problem of re-representing a pattern in another space without loss of information. The Karhunen–Loève expansion for stochastic processes is most commonly used. However, it is rare in pattern recognition that the features that describe most of the pattern are most effective in discriminating between two patterns. Of this, Fukunaga is well aware. Moreover, a mapping from m-space into n-space usually involves measurement of all n of the variables on the pattern itself and thus does not save measurements, but simply computes a hopefully sufficient statistic for the discrimination problem. Fukunaga goes on in Chapter 9 to look at feature selection for two distributions. There he brings into account these distinctions—distinctions which I feel are not given full treatment in Andrews and Meisel. Patrick, on the other hand, has a great deal to offer about all of these subjects.

In Chapter 10, Fukunaga discusses nonlinear mappings. Here, on page 289, another trap opens up. The author says:

> The minimum number of parameters required to account for the observed properties of the data is called the *intrinsic dimensionality* of the data set, or, equivalently, the data-generating process.

This is clearly incorrect, since any finite number of real numbers can be mapped 1-to-1 into the unit interval by interleaving the decimal expansions, thereby reducing a finite-dimensional vector space to a one-dimensional unit interval. Some properties like preservation of linear vector spaces must be posed in order to make this definition meaningful. Perhaps the concept of Hausdorff dimension is what is needed. Fukunaga goes on to discuss some multidimensional scaling procedures in a fairly interesting way and then gives separability enhancement by nonlinear mapping—another interesting topic.

Finally, in Chapter 11, Fukunaga discusses clustering procedures—both parametric and nonparametric. Here the discussion is quite nice.

Despite the cited errors, Fukunaga's book leaves me with a good feeling. Many topics are covered with dispatch. Also the book shows hints of the author's own careful and valuable research. It is too bad that Fukunaga's recent work on error estimation was not ready for inclusion when his book was published.

Fundamentals of Pattern Recognition—Edward A. Patrick.

Patrick's book is almost entirely statistical in nature with an overwhelming emphasis on the Bayesian approach. He explicitly develops the *a posteriori* distribution for many families of distributions and proves the convergence of the risks to zero in those cases where the underlying distributions are identifiable. The novelty of Patrick's approach is that he puts emphasis on the decomposition of mixtures. However, in this he still is less successful, both in developing the subject and in his results, than are Duda and Hart.

The author seems to get sidetracked at times. For example, in Chapter 2, "Maximum *A Posteriori* Estimator Versus Bayes Estimator," he develops some properties of Bayes estimators under squared-error loss criteria. Then in Section 2.9, he goes into a Gaussian example without labeling it as an example; continues on in 2.10 with a continuation of that example where he adds some ad hoc clustering considerations on growing clusters and then goes into an extended discussion of reproducing distributions (primarily of interest with respect to learning or updating schemes for pattern recognition). Then in Section 2.12 he goes on to "Stochastic Approximation." The discussion is very rambling and at no point is it clear that stochastic approximation is what is needed here. For example, when the author considers the decomposition of Gaussian mixtures by observations of random variables drawn according to a mixture distribution, it is clear that a standard maximization over the conditional distribution of the data given the parameter will yield the maximum-likelihood estimate without any recourse to stochastic approximation, because of the known functional form of the distributions. Moreover, no essential characteristic of the stochastic approximation problem is used by Patrick. For example, although Patrick updates his estimates by means of the stochastic approximation algorithm, he does not update the value of the argument at which he will attempt to measure the new value. It is only when the observation random variable depends critically on the choice of the argument that stochastic approximation theory is needed to guarantee convergence. Thus the developed relation between the mixture decomposition problem and stochastic approximation is superficial and misleading.

Abstract Ideas

Patrick's first chapter goes into a number of background concepts in topology and topological spaces and is full of omissions and errors. For example, the author says, "Every continuous function can be approximated by a polynomial, thus the set of polynomials is dense in the set of continuous functions." This example is untrue except for functions restricted to a compact set. For example, e^x does not have an ε-approximation by a polynomial of finite degree except of course when restricted to a compact set. Of course, focusing attention on this section ignores the primary issue of why the section should be included at all. As far as I can see this section is not integrated into the rest of the book and could easily be eliminated. Since the section is so sloppy, its loss would be a gain.

Mixtures

Patrick, as he has frequently pointed out in the literature, was the first to use the idea of mixtures and identifiability of mixtures in the pattern recognition context. These ideas were substantially developed at the time by Teicher and subsequent work was done by Yakowitz and Spragins. The author gives a nice proof of the necessary and sufficient conditions for the identifiability of finite mixtures, namely that the family of distribution functions must be linearly independent over the reals. This was proved by Teicher, and the author follows the proof of Yakowitz and Spragins. This section does not have as satisfactory notation nor is it as clear an exposition as that given in the paper by Yakowitz and Spragins.

Patrick's discussion of distribution-free tolerance regions is quite good. The basic idea is given by the following example. Let $x_1, x_2, \ldots, x_n$ be independent, identically distributed random variables drawn according to some unknown distribution function $F(x)$. Then these points partition the real line into $n + 1$ regions. It can be shown that the probability that a new sample will fall in the ith region is precisely $1/(n + 1)$ for $i = 1,2,\ldots,n + 1$. In many dimensions, the possibilities for creating tolerance regions becomes much greater, and Patrick discusses them fully.

Patrick's presentation of reproducing densities and Bayes estimates on pages 85–107 is quite interesting and utilizes the concept of sufficient statistic in a very nice way to achieve an agreeable and useful generality in the discussion. This work was started by Raiffa and Schlaifer and was introduced into the pattern recognition literature by Spragins.

Probability and Statistics

The errors in Patrick's definitions of convergence are astounding. For example, on pages 56, 57, Patrick says, "or, put in terms of the

relative frequency concept of probability, the latter requirement is that the relative frequency of experiments producing a converging sequence is close to 1." The author goes on to say:

> Convergence w.p.1 will be denoted, for any $\varepsilon > 0$ and $\delta > 0$,
>
> $$p\{\eta: \|\gamma_n - \gamma\| < \varepsilon\} > 1 - \delta, \qquad n > n_0,$$
>
> $$p\{\eta: \lim_{n\to\infty} \gamma_n = \gamma\} = 1$$
>
> . . . where the probability distribution is understood to be over the possible experiments or sequences η, i.e., the number of experiments for which $\|\gamma_n - \gamma\| > \varepsilon$ can be made small.

The verbal description is complete nonsense and shows a lack of understanding of probability theory. The notions of relative frequency and measure are confused. Also, the first definition is convergence in probability rather than convergence with probability 1 and certainly is not equivalent to the condition given in the second equation. As if to emphasize this, the author goes on at the bottom of page 57 to give the definition of convergence in probability by copying precisely the first equation in the aforementioned quotation.

The author considers the concept of consistent estimation with some success but comes to grief in Section 2.4 on sufficient statistics. Here Patrick says:

> The author believes that the concept of sufficiency should be defined as what is sufficient to represent the *a posteriori* density $f(b \mid \dot{x}_n)$ of the parameter vector b characterizing the sample probability density $h(x)$."

This is the restricted Bayesian viewpoint shared in Fukunaga's exposition, but nicely discussed by Duda and Hart. Finally we have, "The following definition of a sufficient statistic is the statistician's definition formulated before mixture densities were studied . . . ." Little need be added to the case made by these quotations other than to remark on the particularly parochial point of view taken by the author that the concept of sufficient statistic is somehow changed when mixture densities are taken into account. Moreover, it is clear that the author is not making effective use of his notation and that despite its complexity, dozens of undefined concepts float though his definitions.

Introduction to Mathematical Techniques in Pattern Recognition—Harry C. Andrews.

Andrews' book is as inviting as Patrick's is formidable. The printing is large and the layout spacious. An idea of the density of the print in Andrews' book can be gotten by noting that there are only 12 references printed per page in the reference section at the end of each chapter. There are 8 pages of references following the first chapter alone. But the content of the book does not do justice to the subject.

Feature Selection

The author makes a convincing case for breaking up the classification problem into a pattern space, feature space, and classification space. He then errs by saying that the classification space is k-dimensional in the k-category problem. What he means to say, of course, is that the classification space has k points.

Andrews' approach to feature selection is almost entirely of the transformation from n-space to m-space variety. This approach largely ignores the fact that the original n measurements have to be made in order to calculate the transformation $y = Ax$. Thus this is not so much feature selection as it is data reduction, and of course in terms of the cardinality of the space, both Euclidean m-space and n-space have the cardinality of the real line.

I like Andrews' inclusion of a discussion of intrinsic dimensionality and multidimensional scaling. This work of Sheppard and Bennett in psychology seems to be very appropriate for the data analysis of pattern recognition. The question is how to take a collection of points in n-space and find the intrinsic dimensionality of it. In other words, how to find a manifold of low dimension, roughly linear, that approximates the points suitably. The discussion here seems thorough and interesting.

In Chapter 3, on "Distribution Free Classification," Andrews follows the classical lines of Nilsson's development in his book, *Learning Machines*, 1965. He discusses linear separability algorithms for determining separability by polynomial discriminant functions. Andrews does not discuss the capacity of a linear threshold device.

I am afraid that Andrews has fallen into the temptation of putting into the book what he knows rather than what he knows that is associated with pattern recognition. It certainly is fair to put in an army of techniques that seem to be related to the problems of statistical pattern processing, such as multivariate analysis, information theory, and so on. But some of it must fall together, some of it must be anchored in the reality of the problem—and in a book, at least, these techniques cannot be allowed to languish in isolation. Otherwise we have a collection of classical topics unchanged by their association with the subject. A list of references would do as well.

For example, Andrews introduces the rate-distortion concept to pattern recognition by saying:

> Appendix B . . . is introduced as a possible tool in the feature selection task. The ideas are purely conjectural but appear well-founded in the information-theoretic context of the book.

In Appendix B, the author says:

> Traditionally the fidelity criterion has been a mean square error estimate or other function estimating type of process typically useful for waveform analysis. However, in the context of the pattern recognition environment, if the fidelity criterion is misclassification error rate, then the rate-distortion function will provide a lower bound on the dimensionality (bits) necessary to achieve a particular distortion (misclassification error).

This certainly does not apply to the two-class problem and the author does not provide any examples where it does apply. For example, in the two-class problem, no matter how many patterns there are and no matter what the nature of the pattern space is, it requires only one bit of information to achieve the minimum classification error. Namely, one needs to know whether a given point does or does not lie in a certain designated subset of the pattern space (for the statistical formulation of the problem, this is simply the partition induced by the Bayes decision rule). Thus when the author goes on to say:

> This, then, may provide a guide to the best feature selection possible for a given misclassification error in terms of dimensionality of bits of encoding (storage). Toward obtaining this bound, the following section investigates the rate-distortion function.

We see that the effort, so directed, will come to naught. In fact rate-distortion theory has to do with a growing number of "quantization points" as the number of samples increases.

In the section on nonparametric classification, Andrews incorrectly states that the Parzen estimator of the density function will converge to zero if the window size $h(n)$ converges to 0 in the limit as n tends to infinity. While that is necessary, it is also necessary that $nh(n)$ tend to infinity in the limit. Thus, not only does Andrews not give the proof of the convergence of the Parzen estimator as is given in Fukunaga, Patrick, and Duda and Hart, but he omits the crucial conditions for the convergence.

In Chapter 5, under the title "Nonsupervised Learning," Andrews treats clustering procedures and considers chaining methods, the potential method, mode estimation, and minimal spanning trees. Inexplicably the author fails to give any properties of the minimal spanning tree. The definition he gives is, "The minimal spanning tree is defined to be that tree or connection of all points in the set to their closest neighbors such that the entire tree is minimum," page 150. He later gives another definition almost as poor. It is well known (Kruskal) that the following algorithm results in a minimal spanning tree for a graph with weighted edges: form a graph consisting of the edge of lowest weight, next lowest weight, etc., rejecting any edges forming loops. This process terminates with a minimal spanning tree.

Andrews goes on to say:

> Because the minimum spanning tree is unique to a set of points in terms of a minimum total weight, it is possible to use the tree as a basis for certain cluster detection techniques combining both distance properties and density properties described below.

Actually, the total weight is unique—but the tree is not. Having all edge weights different guarantees a unique tree, however.

Polynomial Discriminant Functions

Most of the authors have, with the exception of Meisel, in my opinion, incorrectly perceived the strength and weaknesses of the polynomial discriminant function approach of Specht. In this approach, one first approximates the probability density functions for classes 1 and 2. Each probability density is written as the sum of normal distributions centered at the various sample points, the variance of which is some fixed constant σ^2. Obviously one could then find the likelihood ratio with respect to these two empirical density estimates and use the Bayes rule with respect to it. However, this is considered to be too complex in that it uses all of the samples. Specht proposed that each of the normal distributions be expanded in a truncated Taylor series and that the sums of these truncated Taylor series expansions would then be a good estimate of the underlying density. The trouble with all of this is that each truncated Taylor series is good in a region around its particular sample and in fact blows up elsewhere. Thus the sum of these Taylor series expansions will be a poor approximation everywhere. Not only that, but if we fix the polynomial expansion to the degree k, the sum of these polynomial expansions will also have degree k. Thus we find ourselves in the curious position of estimating an underlying density by a polynomial of degree k. But, of course, a polynomial of degree k has an infinite integral that either diverges to plus infinity or minus infinity depending on the oddness or evenness of the degree of the polynomial. It is conceivable, I suppose, for a large enough sample size and for a clever enough choice of the smoothing parameter σ^2, that this polynomial estimate would be a reasonable fit to the density—but there is no general guarantee of this and certainly no proof in the literature.

Andrews looks at the polynomial discriminant function technique of Specht and says on page 129:

> Also by increasing the smoothing parameter σ, the number and order of coefficients necessary for a given accuracy become drastically reduced. As σ becomes small, the true δ-function distribution will result in a point of probability at each prototype. This results in 'nearest neighbor' classification.

The problem with this argument is that as σ goes to 0, the region in which the approximation is good around the given sample point also tends to zero. Thus, while it is true that there is a peak of the polynomial at the data point, it is also true that the function blows up badly far from this point. Thus, instead of getting the nearest-neighbor classification, we will be getting the *farthest*-neighbor classification, because a polynomial tends to infinity for values of x far from its zeroes. None of these complaints would apply if, in fact, the polynomial approximation to the Parzen windows had not been made—because in this case, the component density centered at the sample points tend to 0 in the limit as we retreat from the sample point.

Computer-Oriented Approaches to Pattern Recognition—William S. Meisel.

The contents of this book include statistical and parametric methods, optimization techniques, linear programming, linear discriminant functions, probability density approximations, piecewise linear discriminant functions, cluster analysis and unsupervised learning, and feature selection.

The book is easy to read and has a balanced perspective.

Feature Selection

Meisel gives an interesting discussion of dimensionality and discusses the up-to-date work of Foley in which Foley shows that if m arbitrarily labeled samples are randomly thrown down according to uniform distribution in the unit n-cube, then a certain optimal linear transformation seems to indicate that the categories are widely separated unless m/n is approximately 4 or more. Since the patterns from both sets are drawn according to the same distribution, this would lead the experimenter to a spurious conclusion. However, Meisel then goes on to give an incorrect example. He says:

> Consider, for example, a hypothetical situation: Suppose the designer is presented with a hundred labelled 2-dimensional samples from each of two classes and suppose a linear hyperplane exists which separates the samples with 100% accuracy. Since m/n is 50, the designer may be confident he can achieve a high accuracy on unlabelled samples.

Meisel then adds 98 features to each sample chosen from "a source of random noise." Now m/n is only 1. Then:

> . . . if he proceeds (with a resigned sigh) to obtain a perfect separating hyperplane by methods such as described in Chapter IV, it must intersect the original two-dimensional plane in a two-dimensional hyperplane which separates the two-dimensional samples with 100% accuracy.

This last statement, of course, is false. There is no reason why a separating hyperplane for the 100 samples in 100-dimensional space need separate the projections of these samples in a given 2-dimensional subspace. In fact, it would be very unlikely to have this property. Throughout this discussion, Meisel seems unaware of any work done on the capacity for linear threshold systems. This result says that the probability is precisely one-half that a random dichotomy of $2n$ points in general position in Euclidean n-space is linearly separable. A discussion of intrinsic dimensionality and generalizability should treat this.

Linear Separation

Meisel gives a good discussion of the simplex method approach for finding the best linear separating hyperplane and, in general, finding the best hyperplane when no linear separation is possible. Of the other books, only Duda and Hart treat the simplex method. In general, Meisel's treatment of linear discriminant functions is not as good as Nilsson's original treatment in his book nor does the author adequately acknowledge his debt to Nilsson. Also, as we have mentioned, although algorithms for determining the solution of the system of simultaneous linear inequalities are given, no statements about the probability of the solution are given even though these can be found in Nilsson's treatment.

Nearest-Neighbor Rules

On page 29, Meisel discusses nearest-neighbor classification. He mentions the following objections to it: 1) it assumes that the distance between points is a legitimate measure of the similarity of the patterns they represent, 2) the nearest-neighbor algorithm can be criticized for making poor use of all the information available, and 3) it is best for a small number of samples of reasonably low dimension. However, nowhere does he mention the known nonparametric bounds on the large sample risk of the single nearest-neighbor rule. Most of the more sophisticated procedures that the author discusses in this book cannot be shown to have any properties relative to the Bayes risk. Thus the author's omission is not easily dismissed. Also, convergence of the Fix–Hodges classification procedure is not mentioned.

Density-Function Estimation

The author devotes two chapters to the estimation of probability density functions. He omits showing the convergence of the histogram estimate to the best density approximate of the quantized histogram form even though this proof involves nothing more than the law of large numbers. Next, making the same mistake that Andrews makes in his book, Meisel states that the only condition on the Parzen estimates is that the window size $h(n)$ tends to 0 in the limit as n tends to infinity. While this indeed gives an unbiased estimate of the underlying probability density function, this in no way guarantees that the variance of

the estimate tends to 0. For this we need the additional constraint that $nh(n)$ tends to 0 in the limit as n tends to infinity. Also no proofs are given, although the proof follows in the few lines given Bochner's theorem and, in fact, Parzen proves convergence of his estimate in two pages in the *Ann. Math. Stat.* I seriously disagree with the authors on their omission of proofs. The inadequacy of the conditions in their theorems amply attests to the lack of wisdom in not demonstrating how the conditions relate to the statement of the theorem.

The author gives a thorough treatment of Specht's polynomial discriminant function technique and in my opinion, gives the only adequate warning about the farthest-neighbor rule properties of this procedure. After remarking that the individual expansions of the density kernels are good only in hyperslices through the origin, he goes on to say, "One might ask, however, if the particular approximation chosen to x,y in the above algorithm is the best realization of a good idea." Meisel goes on to modify Specht's definition to given an approximation which is good in the region of interest.

The index for Meisel is not as good, for example, as Nagy's 1968 bibliography for his survey paper on pattern recognition. Nor, for that matter, is Nagy's survey paper listed in the bibliography.

Meisel has an extensive chapter on cluster analysis but does not go into proofs. However, his exposition is much better than Andrews'. For example, on the subject of minimal spanning trees, Meisel says:

> A tree is a connected graph with no closed circuits. A spanning tree of a connected graph G is a tree in G which contains every node of G. The weight of a tree is the sum of the weights of its edges. A minimal spanning tree is a spanning tree of minimal weight.

Inexplicably, Meisel does not offer the easy algorithm for finding the minimal spanning tree.

My overall criticism of Meisel's book is not that he leaves out some topics, but that in the topics that he treats he leaves out the most important parts. Thus the book superficially appears to cover a wide range of interesting topics. However, the author has not mastered the literature on most of them. Nonetheless, the book has an original po int of view and several nice results due to the author including his extension of the polynomial discriminant function approach.

In balance, Meisel's book seems to dominate Andrews' both in the substantial areas in which they overlap (and even agree—many problems are *precisely* the same) and in the point of view and the selection of extra topics.

Conclusions

Pattern Recognition, like Artificial Intelligence and Cybernetics, has caused profound frustration among its devotees, simply because of the softness of the subject and the ambitious goals. The layman continues to shrug at modest progress and will not even be surprised when machines see, play chess, and score baseball games from radio broadcasts. In Williams' "Cat on a Hot Tin Roof," one of the characters drank, he said, until he felt that "click" in the back of his head. Despite much solid progress no one has heard the click yet in pattern recognition, and we are all suffering in the meantime—the authors of these books not the least.

Thomas M. Cover is a Professor in the Departments of Electrical Engineering and Statistics at Stanford University. He has written numerous papers on pattern recognition, including papers on the capacity of perceptrons, bounds on the risks of nearest-neighbor rules, and finite-memory learning theory. He is currently Program Chairman (with A. Rosenfeld and L. Kanal) of the IEEE Conference on Pattern Recognition, Washington, D.C., October 1973, and is the Associate Editor for Pattern Recognition, Hypothesis Testing and Learning for the IEEE Transactions on Information Theory. In 1972 he was President of the Information Theory Group. His current interests in pattern recognition are in the application of Kolmogorov algorithmic complexity to inference procedures for pattern recognition and in the area of learning with finite memory.

BOOKS RECEIVED

George L. Turin

Digital Electronics, F. Dokter and J. Steinhauer (New York: Barnes and Noble, 1973, 270 pp., $19.75). Half of this book in the Philips Technical Library is devoted to the development of the mathematical basis of digital techniques, including lengthy chapters on coding and on switching algebra. The remainder is on circuit realization of logical functions and on solid-state technology.

Dynamo II User's Manual, A. L. Pugh, III (Cambridge, Mass.: MIT Press, 4th ed., 1973, 92 pp., $3.95 ppbk). "Dynamo is a compiler for translating and running continuous models . . . (and) has been designed for the person who is problem-oriented rather than computer-oriented."

Sparse Matrices, R. P. Tewarson (New York: Academic Press, 1973, 160 pp., $11.95). ". . . describes all the significant direct methods of computation involving large sparse matrices . . . (including) . . . the most reliable algorithms for digital computers . . ."