LEARNING IN PATTERN RECOGNITION

Thomas M. Cover [†]

STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Abstract

This paper is specifically concerned with the problem of inferring from a finite set of patterns the classification of an unknown pattern. A discussion of the general problems inherent in the concept of "learning" and "data reduction" are discussed from a standpoint of measurement selection for the general pattern recognition problem. A brief history of the existent work in empirical Bayes and compound sequential Bayes procedures will be presented. It is felt that these procedures are basically non-Bayesian, despite their names, and are therefore especially suited to problems arising in pattern recognition. Finally, a discussion is made of some nonparametric approaches to the problem of the classification of an unknown pattern when the only information on the underlying distributions associated with the various categories is that which can be obtained from a finite number of samples.

I. Patterns and Measurements

I.1 THE PATTERN SPACE

This paper is concerned with the general problem of inferring from a set of patterns the correct classification of a new pattern. No attempt will be made to provide an overall point of view or to provide a complete history of the approaches tangent to those considered herein. I wish simply to present several trains of thought on the subject.

Let $\mathcal{P}$ be an indexed family of patterns $\{P_\alpha\}$. The indexing of the patterns in $\mathcal{P}$ by the parameter α does not imply that $\mathcal{P}$ is countably infinite, nor does it imply that the parameterization is "nice" in the sense that

[†] This work was supported in part by Joint Services Contract NONR--225(83), NR 373 360 and Air Force Contract AFOSR AF49(638)1517.

a being close to a' implies P_a is in some sense close to $P_{a'}$. We mention in passing that the usual distinction between the parametric and nonparametric approaches to statistics (and pattern recognition) lies not in the fact that the set of distributions (or patterns) may be parameterized—since this may always be done—but in the fact that there exists a "nice" parameterization in the parametric case but not in the nonparametric. Thus, in the true parametric case, the parameters should enter significantly into the simplification of the statement of solution of the problem. In the nonparametric case, no parameterization is useful.

The following are examples of possible pattern sets $\mathcal{P}$: the set of all subsets of the plane; the set of all points in Euclidean n-space; the set of all smells, paintings, thoughts, or sensory experiences.

One is tempted immediately to begin to solve the problem, before it is well defined. For example, one might attempt to establish an interesting parameterization on $\mathcal{P}$. Also, it might be interesting to define a "degree of association" of the elements in $\mathcal{P}$ or to define a metric on $\mathcal{P}$. There is certain empirical evidence, especially in the field of art, music, and literature, that humans try blindly to compare and order in an attempt to establish a linear ordering on $\mathcal{P}$. (For what it is worth, those who believe in linearly ordering the arts are termed "absolutists" as opposed to "relativists" in the amusing book "Precious Rubbish".)

We shall be concerned throughout this paper with the following fundamental statement of the pattern recognition problem: Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two indexed families of patterns, and let P be a new pattern. It is desired to associate P with one of the two classes $\mathcal{P}_1$, $\mathcal{P}_2$. On what basis should this assignment be made?

I.2 AN ALGEBRA OF THE SPACES OF MAN'S SENSES

While on the subject of pattern specification, it is appropriate to mention the interesting problem of pattern rerepresentation. I have in mind a systematic description of the algebras of the spaces of man's senses, with special attention to be paid to the formal relationships existing between them. If the rules of combination are well understood in the areas of sight, hearing, smell, taste, and touch, it should be possible to rerepresent the perceptions of one sense in an appropriate form for the perception by another sense. Thus auditory rerepresentations of the visual field for the blind, and visual rerepresentations of sounds for the deaf could be found. It is important, however, that any such rerepresentation be a homomorphism in the sense that the rerepresentation of the superposition of several perceptual experiences in one domain should yield the superposition of the rerepresentations in the other domain. If one object obscures another in the visual field we might like the corresponding sound to "obscure" the other sound in the auditory field. Similarly, complementary images should yield complementary sounds.

Consider for a moment the pattern spaces of sight and sound. Both the ear and the eye are affected pleasurably by pure tones and pure light frequencies respectively. However, the ear has the capability of distinguishing the basic frequency components present in the superposition of a small number of pure audio frequencies. The eye, on the other hand, receives the subjective impression of another pure frequency when presented with the superposition of two pure frequencies. To confuse the picture even further, we note the lack of a corresponding concept in audition to the concept of the color wheel in vision—there are no "primary" audio tones from which the others may be generated.

From these well-known examples and others it may be seen that any attempt to rerepresent one space in terms of another must face the fact that most simple 1-to-1 correspondences will fail to have consistent rules of combination and therefore will not be an adequate sensory rerepresentation.

I.3 SUFFICIENCY IN MEASUREMENTS

In statistics the concept of a sufficient statistic is instrumental in reducing the order of complexity of calculations and the conceptualization involved in hypothesis-testing and estimation problems. For example, $\overline{x}_n = (1/n)\ \Sigma x_i$ is a sufficient statistic for the estimation of the mean of n independent identically distributed random variables $x_1, x_2, \ldots, x_n$ drawn according to a normal distribution with unknown mean and known variance.

Similar considerations may be fruitful in the general area of data reduction, measurement selection, and preprocessing for pattern recognition. Let us introduce the following ideas. Let M be a measurement function, defined on $\mathcal{P}$, taking values in an arbitrary space.

i) Sufficiency with respect to $\mathcal{P}$: M is said to be *sufficient with respect to* $\mathcal{P}$ if M is 1:1. Of course this definition is trivial in the sense that no data reduction is accomplished. M merely rerepresents the patterns. However, several interesting examples of measurements are in this category.

Example. Let M assign to every function P(t), $-\infty \leqslant t \leqslant \infty$, defined on the real line the values that P(t) takes at the points $t = 0, \pm 1/2W, \pm 2/2W, \ldots$ By the Nyquist sampling theorem, if $\mathcal{P}$ is restricted to the set of all band-limited functions with bandwidth W, then M is 1:1.

Example. Let M assign to every subset P of 3-space the set of all its shadow appropriately indexed by the planes upon which the shadows are cast. Then if $\mathcal{P}$ is the set of all convex sets, M is 1-to-1.

Example. Let M assign to every object in 3-space the set of all X-ray photographs of the object appropriately indexed by the directions in which the photographs were taken. Let $\mathcal{P}$ be the set of all everywhere-nonzero X-ray transmissivity functions defined on E^3. Then M is 1:1. It is an interesting exercise to derive an explicit reconstruction scheme for P from its X-rays.

Example. Let M map the horizontal-vertical axes of the visual field into a frequency-time audio signal. (Imagine a black and white scene being played by a player piano or harmonica.) Thus the visual field is rerepresented as a tapestry of sound for the ear. A device implementing this mapping was invented in 1914 under the name "Optophone". Unfortunately, as indicated in the previous section, although the mapping is 1-to-1, the ear does not resolve sounds in the same way the eye resolves visual fields.

ii) Sufficiency with respect to Classification: M is said to be *sufficient with respect to the classification* problem with classes $\mathcal{P}_1$, $\mathcal{P}_2$ if $M(P_1) \neq M(P_2)$, for all $P_1 \in \mathcal{P}_1$, $P_2 \in \mathcal{P}_2$.

Example. Size and shape may be ignored in distinguishing blue objects from red.

Example. Translation, rotation, and scale may be ignored in distinguishing circles from squares.

iii) Sufficiency with respect to an intermediate operation: Let ϕ map $\mathcal{P}$ into another pattern set $\mathcal{P}'$. Then M is sufficient with respect to $(\phi,\mathcal{P})$ if M is sufficient with respect to $\mathcal{P}'$.

Again, this is a trivial definition in the sense that it is merely definition *(i)* with respect to the new set $\mathcal{P}'$. We had the following example in mind:

Example. Let $\mathcal{P}$ be the set of all waveforms. Let $\phi_1(\mathcal{P}) = \mathcal{P}'$ be the set of all waveforms distinguishable in a certain sound conducting medium. Let $\phi_2(\mathcal{P}') = \mathcal{P}''$ be the set of all sound waveforms distinguishable by the human ear. Let M assign every waveform P a canonical waveform with arbitrary preassigned phase relationships among the component frequencies. Then certain physiological evidence indicates that M is sufficient with respect to $(\phi_2 \circ \phi_1, \mathcal{P})$.

iv) Sufficiency with respect to a family of decision procedures: Let G = $\{g_\beta\}$ be a family of decision mappings g_β: $\mathcal{P} \rightarrow \{1,2\}$. *Then* M *is sufficient with respect to* G *and* $\mathcal{P}$ if, for every α, $M(P_\alpha) = M(P_{\alpha}') \Rightarrow g(P_\alpha) = g(P_{\alpha}')$ for all $g \in G$.

Example. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two indexed sets of points in E^n. If G is the set of all decision procedures of the hyperplane type (in which a hyperplane is passed through E^n in such a manner as to minimize the number of errors of misclassification of $\mathcal{P}_1$ and $\mathcal{P}_2$), then the indices of the patterns in $\mathcal{P}_1$ and $\mathcal{P}_2$ may be forgotten—because a hyperplane rule is "symmetric in the data".

v) Statistical Sufficiency: Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two pattern sets. Let the random pattern $\tilde{P}$ be drawn according to a probability distribution f_1 if the underlying pattern $P \in \mathcal{P}_1$ and according to f_2 if $P \in \mathcal{P}_2$, where f_1 and f_2 are defined on a pattern space $\tilde{\mathcal{P}}$. Then $\underset{\sim}{M}$ *is sufficient with respect to* (f_1, f_2) if there exists a factorization $f_i(\tilde{P}) = g_i(M(\tilde{P}))h(\tilde{P})$, i = 1, 2, where the first factor may depend on i but depends on $\tilde{P}$ only through $M(\tilde{P})$ while the second factor is independent of i. This is the well-known factorization criterion [see, for example, Lehman, E.L., "Testing of Statistical Hypotheses", Wiley, 1959, pp. 17–21.]

Example. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be the single point sets {+1} and {−1} respectively, and let $P = (\tilde{P}_1, \tilde{P}_2, \ldots, \tilde{P}_n)$ be a vector of n independent identically distributed observations on the underlying pattern P, where $\tilde{P}_i$ is drawn according to a normal distribution $N(\mu,\sigma^2)$ with known variance σ^2 and mean $\mu = +1$, if $P \in \mathcal{P}_1$, and mean $\mu = -1$, if $P \in \mathcal{P}_2$. The factorization theorem may be used to show that the sample mean

$$M(P) = \frac{1}{n} \sum_{i=1}^{n} P_i$$

is sufficient with respect to f_1, f_2.

II. Data Reduction and Finite Memory

II.1 INTRODUCTORY REMARKS

A great amount of attention seems to have been paid to the general problem of data reduction from the standpoint of reducing the demands on the memory required for a particular data processing scheme. Those data reduction schemes which focus on the existence of many-to-one information-lossless maps and sufficient statistics have only a superficial application to memory reduction, as we shall indicate in the next section. In this chapter we shall mention a new approach, which attacks the memory reduction problem directly. The proofs of Theorems 1 and 2 will appear in [1, 2]. A time-varying solution to the famous two-armed bandit problem with finite memory [3,4] will appear in [5].

II.2 SUFFICIENT STATISTICS AND FINITE MEMORY

Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables (i.i.d. r.v.'s) drawn according to some unknown probability density f(x) defined on the real line. Throughout this section we shall be interested in the hypothesis test $H_0 : f = f_0$ vs. $H_1 : f = f_1$. For a given decision procedure, which assigns each possible observations $(x_1, x_2, \ldots, x_n)$, n = 0, 1, 2, . . . , to H_0 or H_1, we may define $\alpha_n = \Pr\{\text{Decide } H_1 | H_0\}$ and $\beta_n = \Pr\{\text{Decide } H_0 | H_1\}$. Thus α_n and β_n are the probabilities of error of each kind, based on the first n observations, for the given decision procedure.

It is well known that the standard likelihood-ratio decision procedure results in $\alpha_n \to 0$ and $\beta_n \to 0$ exponentially in n, with rates which depend on an information distance between f_0 and f_1. To apply this procedure at time n requires a memory capacity sufficient to store the observations $x_1, x_2, \ldots, x_n$. Thus, even in the simplest case, the memory must grow indefinitely with time. Any truncation of memory to the last k observations, for example—as in the most familiar definition of finite memory [see 3,4]—will preclude the convergence of α_n and β_n to 0, except in the singular case.

It has been frequently observed that the data may be reduced by a sufficient statistic without loss of information. For example,

$$\overline{x}_n = (1/n) \sum_{i=1}^{n} x_i$$

is a sufficient statistic for testing the mean of a normal distribution. However, while it is true the mapping corresponding to a sufficient statistic is many-to-one, and is in this sense data reducing, it is generally not true that the cardinality of the required memory is reduced. For example, in the case of the univariate normal, the mapping from $(x_1, x_2, \ldots, x_n) \in R^n$ to $\overline{x}_n \in R$ leaves the memory requirement uncountably infinite. A possible alternative information-lossless mapping from R^n to R is given by the simple trick of interleaving the digits in the decimal expansions of $x_1, x_2, \ldots, x_n$ to form a single real number. (The possible objection that this mapping lacks the continuity possessed by $\overline{x}_n$ will be obviated shortly.) Clearly, then, if one can store one real number, one may store any finite number of real numbers. We conclude that the statistic $\overline{x}_n$ has not decreased the memory requirement at all.

Particular attention in the work of Dynkin, Spragins, and Abramson [6,7,8] has been paid to the existence of finite-dimensional sufficient statistics (such as $\overline{x}_n$ for the normal). Here, by implication, it would seem that the memory is bounded in some sense by the dimension of the minimal sufficient statistic. Again, from the standpoint of memory capacity, there is no resultant saving in memory. The previously mentioned interleaving decimal expansion yields a mapping of an arbitrary number of univariate observations into a 1-dimensional sufficient statistic, thus accomplishing the same task as the perhaps nonexistent finite-dimensional sufficient statistic.

Even the nice continuity properties of the usual minimal sufficient statistic are not difficult to duplicate. The work of Denny [9,10] establishes the existence of a 1-1 uniformly continuous map of R^n into R, excluding a set of Lebesgue measure zero. If used as a statistic, small errors in memory would be reflected in uniformly small errors in the reconstruction of the data, and conversely. Thus there always exists a continuous 1-dimensional sufficient statistic with respect to $\{f_\theta\}$.

A first step toward defining a statistic with a realistic memory constraint might be to consider rounding off the statistic at each stage. Hopefully, the infinite-accuracy theory would apply directly, and α_n, β_n would

still tend to 0, although at slower rates. Even this is not the case, as the following simple example will show: Let $X_1, X_2, \ldots$ be i.i.d. r.v.'s drawn according to a normal $N(\mu,\sigma^2)$ distribution with unknown mean μ and known variance σ^2. We wish to test $\mu = 1$ vs. $\mu = -1$. Observe that the new statistic $\overline{x}_{n+1}$ may be expressed in terms of the old statistic $\overline{x}_n$ and the current observation x_{n+1} in the form

$$\overline{x}_{n+1} = \frac{n}{n+1} \overline{x}_n + \frac{1}{n+1} x_{n+1}$$

Suppose now that $\overline{x}_n$ may be recalled only to some arbitrary decimal place accuracy. Let $[\overline{x}_n]$ denote the rounded-off version of $\overline{x}_n$. Rounding off at each stage results in the algorithm

$$[\overline{x}_{n+1}] = [\frac{n}{n+1} [\overline{x}_n] + \frac{1}{n+1} x_{n+1}] .$$

To what random variable does $[\overline{x}_n]$ converge? The best hope is that $[\overline{x}_n] \to [\mu]$ wpl. Thus the decision procedure that decides $\mu = \pm 1$ accordingly as $[\overline{x}_n] \gtrless 0$ would result in $\alpha_n, \beta_n \to 0$. Instead, the worst possible situation occurs. As we have shown in [1], $[\overline{x}_n]$ converges wpl to a random variable which has strictly positive probability mass on each of the countably infinite number of lattice roundoff values. In particular, there is positive mass on both sides of the origin. Consequently, α_n, β_n converge to nonzero limits. So the first realistic approximation to the data reduction problem will not resolve the hypotheses. This is true despite the fact that we have used an eminently reasonable procedure and a countably infinite memory. Apparently the memory-constrained hypothesis testing problem must be approached from first principles.

To our surprise, we find that this same problem may be solved with a memory of only 2 states (1 bit). Consider the sequence of statistics $\{T_n\}_1^\infty$, $T_n \in \{-1,1\}$, defined recursively by

$$T_n = \begin{cases} 1, & x_n > \sqrt{2 \log n} \\ -1, & x_n < -\sqrt{2 \log n} \\ T_{n-1}, & \text{otherwise} \end{cases}$$

$$T_0 \ \text{arbitrary} \in \{-1,1\}.$$

We make use of the fact that, for $X_1, X_2, \ldots$ i.i.d. $\sim N(\mu,\sigma^2)$, $\max\{X_1, X_2, \ldots, X_n\} - \sqrt{2\sigma^2 \log n} \to \mu$, in probability. As a consequence of this and the Borel zero-one law, for $\mu > 0$, the event $X_n > \sqrt{2\sigma^2 \log n}$ occurs infinitely often, while the event $X_n < -\sqrt{2\sigma^2 \log n}$

occurs only finitely often, wpl. A corresponding statement holds for $\mu < 0$. Therefore $T_n \to 1$ or -1 accordingly as $\mu > 0$ or $\mu < 0$. We have thus furnished an example of a 2-state memory which resolves the composite hypothesis testing problem $\mu > 0$ vs. $\mu < 0$ with probabilities of error α_n, $\beta_n \to 0$. We generalize this example to arbitrary distributions in the next section.

To summarize our point of view, we admit the utility of sufficient statistics, finite-dimensional or otherwise, in computation, but doubt their utility in the problem of data reduction. First, for multivariate data, there exist trivial data preserving mappings into the unit interval. Denny's work provides uniformly continuous such mappings. Second, the straightforward sequential rounding off of sufficient statistics generally fails to yield α_n, $\beta_n \to 0$. Thus a memory constraint requires a more careful approach than the simple rounding off of the infinite-memory procedures. Third, we might add that the existence of a finite dimensional sufficient statistic is destroyed by a slight distortion of the distribution (without greatly affecting the resolving power of the old statistic). In this sense, the existence of a finite-dimensional sufficient statistic is a "measure-zero" phenomenon, not to be taken too seriously.

In the next sections we shall provide the beginnings of a theory of hypothesis testing with finite statistics in which the hypotheses are learned despite nontrivial data reduction at each stage. Specifically we shall demonstrate the existence of a four-valued (and in some cases two-valued) sequentially updated statistic achieving $\alpha_n \to 0$ and $\beta_n \to 0$. Sole concern will be with the algorithm

$$T_{n+1} = f_n(T_n, x_{n+1}) \, ,$$

where the memory (or statistic) T_n takes values in the set $\{1, 2, \ldots, m\}$. Thus the new value of T depends explicitly only on the old value of T and the current observation.

II.3 LEARNING WITH FINITE MEMORY

Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables drawn according to a probability density function f(x). Consider the hypothesis test $H_0 : f = f_0$ vs. $H_1 : f = f_1$. Define the likelihood ratio $\ell(x) = f_1(x)/f_0(x)$. Let us consider f_i and f_0 such that ℓ is unbounded above and unbounded away from zero, with probability one, under each hypothesis. We shall refer to this as the unbounded likelihood ratio case. The proof of the following theorem is given in Cover [1].

Theorem 1: In the unbounded likelihood ratio case, there exist sequences of thresholds $\{\bar{\ell}_n\}$, $\{\underline{\ell}_n\}$ such that the algorithm

$$T_{n+1} = \begin{cases} 1, & \ell(x_n) > \bar{\ell}_n \\ -1, & \ell(x_n) < \underline{\ell}_n \\ T_n \, , & \text{Otherwise} \end{cases}$$

results in $T_n \to 1$ wpl under H_1 and $T_n \to -1$ under H_0. Thus $\alpha_n \to 0$ and $\beta_n \to 0$ with a 2-state memory, under either hypothesis. With probability one, only a finite number of mistakes will be made by $\{T_n\}$.

There are certain heuristic considerations which make it appear unlikely that learning is possible in the bounded likelihood ratio case. In the Bayesian formulation, for example, where prior probabilities are associated with the two hypotheses, the rule which stores the Bayes decision at each stage will not learn. Eventually the posterior probabilites will be such that no single observation will yield a change in decision.

Fortunately, experiments of arbitrarily large information may be compounded from experiments of bounded information by the artifice of looking for sequences of events before changing the state of the memory. (This point of view yields some interesting comments [5] on the 2-armed bandit problem with finite memory [3].)

Consider the basic problem of testing the compound hypothesis that a coin with bias p has bias $p \geqslant p_0$ vs. the compound hypothesis that $p < p_0$. Note that the general two-hypothesis testing problem with $\{\tilde{X}_1\}$ i.i.d. may be put in this framework under the correspondence

$$X_i = \begin{cases} 1, & \ell(\tilde{X}_i) \geqslant 1 \\ 0, & \ell(\tilde{X}_i) < 1 \end{cases}$$

and

$$P_0 = \frac{1}{2}(\Pr\{X_i = 1 | H_1\} + \Pr\{X_i = 1 | H_0\}) \, .$$

Theorem 2: Let $X_1, X_2, \ldots$ be a sequence of i.i.d. Bernoulli r.v.'s with $\Pr\{X_i = 1\} = p$. There exists an algorithm with a 4-state memory for which the hypothesis $p \geqslant p_0$ vs. $p < p_0$ is resolved with limiting probability of error zero under either hypothesis. (Proof given in [1].)

Example 1: Let $X_1, X_2, \ldots$ be i.i.d. real valued normal random variables with mean zero and unknown variance σ^2. Let

$$Y_i = \begin{cases} 1, X_i^2 > c^2 \\ 0, X_i^2 \leqslant c^2 \end{cases} \, .$$

Note that $\Pr\{Y_i = 1 | \sigma^2\} = 2\Phi(c/\sigma)$. Let $p_0 = 2\Phi(c/\sigma)$. Then the 4-state test described in this section will test $\sigma^2 \geqslant c^2$ versus $\sigma^2 < c^2$, with limiting probability of error zero. In the final analysis, we are testing the hypothesis $\Pr\{X \in X_1\} \geqslant p_o$ vs. $\Pr\{X \in X_1\} < p_o$.

Example 2: In the univariate case, let

$$Y_i = \begin{cases} 1, & X_i \geqslant c \\ 0, & X_i < c\,. \end{cases}$$

The test in this section resolves $F(c) \gtreqless p_o$, where F is the unknown cdf of x. If $p_o = 1/2$, we have a nonparametric finite-memory test of whether or not the median is greater than c.

II.4 COMMENTS ON THE ALGORITHM

In an attempt to solve the hypothesis-testing problem under realistic memory constraints, we have investigated the straightforward quantization of infinite-memory schemes based on sufficient statistics. Somewhat to our surprise we found that in general such schemes will not learn (i.e. α_n and β_n are bounded away from 0.) However, a different approach establishes that a four-state memory is sufficient for learning under any two hypotheses.

We should remark that the hypothesis-testing problem is essentially infinite in the sense that no finite number of samples yields $\alpha_n = 0$, $\beta_n = 0$, except in the singular case. Thus some capability must be retained for accepting this infinity. The variation of f with n is a natural way to meet this requirement. Thus the algorithm has been factored into two parts. That dealing with the data is finite, while the part concerned with the data processing is unbounded. The important point is that the data processing algorithm $f_n(\cdot)$ is specified independently of the data. We do not cheat by storing data in the description of f_n (See further discussion in Reference 5.)

There is a nice complementary between the theory of Turing machines and that of hypothesis testing with finite statistics. Turing machines have essentially infinite memory (an infinite tape) with finite computation (i.e., an f_n which is independent of n), while the finite-memory algorithm has finite memory and unbounded complexity of computation f_n. Each problem seems to be stated in its most natural form. Bounding the length of the tape for a Turing machine results in a hopelessly combinatorial theory. Similarly, for finite-memory hypothesis testing algorithms, allowing the memory to be infinite yields the theoretical trivial classical hypothesis-testing problem, while restricting f_n to be independent of n yields a hopelessly combinatorial theory which, in any case, precludes the convergence of α_n and β_n to 0.

III. Nonparametric Pattern Recognition

III.1 CHOOSING THE BEST NONPARAMETRIC CLASSIFICATION RULES

Suppose that n labeled observations $(x_1,\theta_1), (x_2,\theta_2), \ldots, (x_n,\theta_n)$ are presented and a new pattern (or observation) x_{n+1} is to be classified. Here we assume that the x_i's take values in some observation space and that each θ_i equals 0 or 1. Let $g(x_{n+1}; (x_1,\theta_1), \ldots, (x_n,\theta_n))$ be an arbitrary estimator defined on $X \times (X \times \Theta)^n$ which assigns to every x_{n+1} an estimate $g = 0$ or 1 based on the n training samples. Thus g induces a partition of X into 2 sets, this partition partaking of the randomness of the training set $\{(x_i,\theta_i); i = 1, 2, \ldots, n\}$. Let $G = \{g_\alpha\}$ be a set of such decision rules. For example, G might be the set of all k-nearest-neighbor procedures, for all k, together with all procedures of the separating hyperplane type [11, 12,13]. How is one to select the "best" procedure for the assignment of x_{n+1}?

Either of the following basic assumptions about the homogeneity of $\{(x_1,\theta_1), \ldots, (x_n,\theta_n), (x_n,\theta_{n+1})\}$ will suffice for our further remarks:

i) $(x_1,\theta_1), \ldots, (x_{n+1},\theta_{n+1})$ is a collection of n+1 independent identically distributed random variables. Dependence between x and θ is allowed.

ii) $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n, \tilde{x}_{n+1} \in X$ and $\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_n, \tilde{\theta}_{n+1} \in \Theta$ are arbitrary sequences. A permutation π of $\{1, 2, \ldots, n+1\}$ is chosen at random according to a uniform distribution on the set of (n+1)! permutations. Then the assignment $x_i = \tilde{x}_{\pi(i)}$; $\theta_i = \tilde{\theta}_{\pi(i)}$; $i = 1, 2, \ldots, n+1$, is made.

Let $S(\hat{g})$ be the proportion of incorrect assignments obtained by $\hat{g}$ when $\hat{g}$ is used to assign each of the x_i's, $i = 1, 2, \ldots, n$ in terms of the remaining x_i's. To be precise, let σ be a permutation of $\{1, 2, \ldots, n\}$. Let $\delta(\theta,\hat{\theta}) = 1$ or 0 accordingly as $\hat{\theta} \neq \theta$ or $\hat{\theta} = \theta$. Then $S(\hat{g}) = 1/n! \sum_{\sigma} \delta[\theta_{\sigma(i)} ; \hat{g}(x_{\sigma(i)} ; (x_{\sigma(j)}, \theta_{\sigma(j)}, j = 1, 2, \ldots, n, j \neq i)]$. Observe that $0 \leqslant S(\hat{g}) \leqslant 1$. Now, for a given $\hat{g}$, $S(\hat{g})$ will be a random variable, the distribution of which is governed by the distribution of the (x_i,θ_i)'s.

It is understood that $\hat{g}$ will be used to classify x_{n+1} in the following randomized way. First a permutation σ of $\{1, 2, \ldots, n\}$ will be chosen according to an equiprobable distribution on the n! permutations, and then x_{n+1} will be given the classification $\hat{\theta}_{n+1} = \hat{g}(x_{n+1} ; (x_{\sigma(1)}, \theta_{\sigma(1)}), \ldots, (x_{\sigma(n-1)}, \theta_{\sigma(n-1)}))$. The permutation σ plays the role of symmetrizing the data so that the order in which the observations occur is irrelevant. Clearly this precaution is unnecessary under assumption *(i)*. Note also that the $(x_{\sigma(n)}, \theta_{\sigma(n)})$ sample is deleted so that $\hat{g}$ is defined on the same number of variables for which it was evaluated. We define the risk, or probability of error, of this procedure to be $R(\hat{g})$ where

$$R(\hat{g}) = \Pr\{\hat{\theta}_{n+1} \neq \theta_{n+1}\},$$

where the probability is taken with respect to the distribution of the (x_i,θ_i)'s (under either assumption *i)* or *ii)*) as well as the distribution on σ.

Now we note the interesting fact that $S(\hat{g})$ is an unbiased estimate of the probability of error in using $\hat{g}$ on x_{n+1} in the sense that $E\{S(\hat{g}) | R(\hat{g})\} = R(\hat{g})$ where again the expectation is taken over the distribution on $(x_1, \theta_1), \ldots, (x_n,\theta_n)$ and σ. We shall not give the simple proof here.

The optimal classifier in $G = \{g_\alpha\}$ is the one which minimizes $R(g_\alpha)$. Since, in the nonparametric case, $R(g_\alpha)$ is unknown, we propose to choose the classifier which minimizes $S(g_\alpha)$. The hazards are obvious—comparing means on the basis of the comparison of random variables having the corresponding means may yield arbitrarily poor results, depending on the distributions. In spite of this, I feel that this procedure will result in good decision rules in practice. The next section illustrates some of the families of decision rules which we have in mind for G.

It may be worth noting that with high probability the best procedure in G may be chosen if one is willing to discard a large portion of the samples. Let the n samples in the training set be divided into k disjoint subsets each containing r samples. Let g be defined on $X \times (X \times \Theta)^r$. Then g will achieve scores $S_1(g), S_2(g), \ldots, S_k(g)$ on the corresponding blocks of data. Under assumption *(i)* the blocks are independent. Thus $S_i(g)$, $i = 1, 2, \ldots, k$, is a set of independent identically distributed random variables with common mean $R(g)$. Hence

$$S = \sum_{i=1}^{k} S_i(g)/k$$

is an unbiased estimate of $R(g)$ for which the variance tends to zero at the rate $0(1/k)$.

Notice that G dichotomizes the family of all distributions $\{f_\alpha(x,\theta)\}$ into sets for which G does and does not perform well. This point of view complements the usual point of view in which a nonparametric family of probability distributions $\{f_\alpha(x,\theta)\}$ is presented and one is asked to find a set G of reasonable decision procedures with respect to it. Here, we specify G first and then ask for the natural family (or families) of distributions for which G is best suited.

III.2 NEAREST NEIGHBOR ESTIMATE OF THE BAYES RISK

Let $(x_1,\theta_1), (x_2,\theta_2), \ldots, (x_n,\theta_n)$ be a set of n independent identically distributed random variables drawn according to a joint probability density $f(x,\theta)$. Let $\Theta = \{1,2\}$. If f is known, the minimum risk incurred in estimating the unknown parameter θ associated with a random variable observation x is defined to be the Bayes risk R^* (see Section IV.2). Now suppose that we are in the nonparametric case in which the only knowledge of f is that which may be inferred from the n samples. In the previous section we have asked how to classify a new sample. The question we ask here is "What is R^*?".

We wish to propose a "deleted nearest-neighbor" rule as a measure of the *intrinsic complexity* or *interleavedness* of the data in the classification problem. As a biproduct, an estimate of the Bayes risk will be obtained.

To be specific, let R_{n-1} denote the probability that the nearest neighbor to the i^{th} observation x_i has a different classification from x_i, where the probability is defined with respect to the overall distribution of $\{(x_i,\theta_i)\}_1^n$. (Since the (x_i,θ_i)'s are independent identically distributed, R_n is independent of i.) Let S_{n-1} = [number of indices $i = 1, 2, \ldots, n$ for which the classification of the nearest neighbor to x_i disagrees with θ_i]/n. Since S_{n-1} may be considered the sum of identically distributed 0-1 random variables each with mean R_{n-1}, we see that

$$E\{S_{n-1}\} = R_{n-1} \; .$$

Now, R_{n-1} is seen to be, by definition, the (n−1)-sample nearest-neighbor risk investigated in Cover and Hart [11]. In that investigation it was proved that $R_n \rightarrow R_\infty$, almost without restriction on f, where $R^* \leqslant R_\infty \leqslant 2R^*(1-R^*)$. It is further shown in Cover [14] that, under slight further assumptions on the allowable family of f's, $R_n \leqslant R_\infty + c/n^2$, where c is some constant depending on f. Thus we see that the score S_{n-1} of the NN rule is a random variable having the property that

$$E\{S_{n-1}\} = R_{n-1}$$

and

$$R^* \leqslant E\{S_{n-1}\} \leqslant 2R^*(1-R^*) + c/n^2 \leqslant 2R^* + c/n^2 .$$

Inverting the two most extreme inequalities we have

$$\frac{1}{2}(E\{S_{n-1}\} - c/n^2) \leqslant R^* \leqslant E\{S_{n-1}\} .$$

Thus R^* is bounded in terms of $E\{S_{n-1}\}$. If it is also true that $S_{n-1} \rightarrow E\{S_{n-1}\}$ in probability, as $n \rightarrow \infty$, which seems reasonable, then S_{n-1} is more firmly established as a bound on R^* in the sense that

$$\Pr\{S_n/2 \leqslant R^* \leqslant S_n\} \rightarrow 1$$

as $n \rightarrow \infty$. Thus we claim that the interval $[S_n/2, S_n]$ is a good estimate of R^*.

For ease of explication, the upper bound $2R^*(1-R^*)$ was replaced by the weaker bound $2R^*$ in the inversion of inequalities. So all the preceding remarks hold for the tighter interval bound $[(1-\sqrt{1-2S_n})/2, S_n]$.

The idea that the NN risk may be used to bound the Bayes risk seems to have been arrived at independently by Cover, Hart, and Whitney[15].

Whether or not it is felt that n is sufficiently large for S_{n-1} to be interesting bound on R^*, the score S_{n-1} yields a great amount of information about the interleavedness or intrinsic complexity of the data. In fact, a large difference between S_{n-1} and its limit R_∞ reflects a small sample size with respect to $f(x,\theta)$. Although S_{n-1} is, in this case, a poor estimate of R_∞, it is probably a very good estimate of the best that any nonparametric decision rule may do in terms of the small sample $\{(x_i,\theta_i)\}_1^n$. In other words, we feel that the failure of the NN rule score to be near its limit is a good indication that every other decision rule based on the n samples will also be doomed to poor behavior. A small sample with respect to the NN rule is probably a smaller sample with respect to more complicated data processing rules.

IV. Asymptotically Bayes Non-Bayesian Procedures

IV.1 PRELIMINARY REMARKS

It is clear in many pattern recognition problems that elaborate statistical assumptions are not justified. Certainly this is the case in character recognition problems. But it is equally clear that many procedures from the realm of classical statistical hypothesis testing are eminently reasonable. The procedures arising in Bayes decision theory, in particular, fall in this class. The reasonableness of these procedures suggests that they may be applicable in a wide variety of problems in which the standard statistical assumptions do not hold or do not make sense.

The purpose of the next few sections is to study statistical decision procedures which are variously known, in the engineering literature, as learning without a teacher, and nonsupervised identification, and, in the statistics literature, as compound and sequential compound Bayes decision procedures. It will be shown that, with a little bit of care (sometimes involving the use of randomized decision procedures), Bayes-like decision rules are asymptotically Bayes even in problems in which the concept of an underlying prior distribution makes no sense. However, we wish to first to review the existent Bayes formulation of the sequential decision problem.

IV.2 THE BAYESIAN FORMULATION

Let $\Theta = \{1,2\}$ be a parameter space with two elements. Let X be an observation space or pattern space. Consider the pattern recognition problem in which the random variable θ is drawn according to the distribution

$$\theta = \begin{cases} 1, \text{ with probability } \eta \\ 2, \text{ with probability } 1-\eta . \end{cases}$$

Suppose that a random observation X is drawn according to the probability density function $f_1(x)$, if $\theta = 1$, and $f_2(x)$, if $\theta = 2$, where f_1 and f_2 are defined on X. Let $\delta(x)$ be a decision function defined on X taking values in $\{1,2\}$. Then, the probability of error R of the decision rule δ is simply given by

$$R = \eta \int (1-\delta(x))f_1(x)dx + (1-\eta)\int \delta(x)f_2(x)dx.$$

The minimum value R^* of R, over all decision rules δ, is achieved by the decision rule

$$\delta^*(x) = \begin{cases} 1, & \Pr\{\theta=1 \mid X=x\} \geqslant \Pr\{\theta=2 \mid X=x\} \\ 2, & \Pr\{\theta=1 \mid X=x\} > \Pr\{\theta=2 \mid X=x\} \end{cases}$$

where $\Pr\{\theta=1 \mid X=x\} = \eta f_1(x)/(\eta f_1(x) + (1-\eta)f_2(x))$ (by application of Bayes rule). The resulting risk R^* is then $R^* = R^*(\eta, f_1, f_2) = \int \min\{\eta f_1(x), (1-\eta)f_2(x)\}dx$. R^* and δ^* are termed the Bayes risk and Bayes decision rule respectively. We shall frequently draw attention to the dependence of R^* on η, f_1, and f_2 by explicitly writing $R^* = R^*(\eta, f_1, f_2)$ as we have in the last equation.

IV.3 SEQUENTIAL BAYES LEARNING

We shall attempt to review some of the work on the general problem of learning in the classification problem. Let $\underline{\theta}_n = (\theta_1, \theta_2, \ldots, \theta_n)$ be an arbitrary sequence of integers $\theta_i \in \{1, 2, \ldots, k\}$, where θ_i denotes the category of the ith sample. Let $\underline{x}_n = (x_1, x_2, \ldots, x_n)$ be the resulting sequence of random observations, where x_i is drawn according to probability density function $f_{\theta_i}(x)$. It is assumed that the x_i's are conditionally independently distributed (conditioned on the θ_i's). Let $\underline{f} = (f_1, f_2, \ldots, f_k)$ denote the set of k probability density functions on X. We shall be concerned with classifying the x's given various states of knowledge about $\underline{x}$, $\underline{\theta}$, and $\underline{f}$.

Frequently we shall not know the f_i's precisely but only that they belong to some set F. When there exists no "interesting" parameterization of F the problem lies in the domain of nonparametric statistics. Let us define

$$\overline{L}_n = (1/n) \sum_{i=1}^{n} L_i$$

to be the random variable losses incurred by a given decision rule $\delta(x)$. A 0-1 loss function (i.e., a probability of error loss function) will be assumed. Let R_n be the expected value of $\overline{L}_n$, where the expectation is taken with respect to the distribution on $\underline{\theta}$, $\underline{x}$, $\underline{f}$, and δ.

The earliest, and one of the most ambitious, treatments in the engineering literature of the unsupervised learning problem is given in Cooper and Cooper [16]. They consider the problem in which the θ_i's are independent identically distributed Bernoulli random variables with $\Pr\{\theta_i = 1\} = \Pr\{\theta_i = 2\} = 1/2$. It is assumed that the functional forms of the f_i's are known up to some unknown parameters. The authors show that the classification problem has a solution if the two equiprobable distributions differ only in location parameters. They exhibit a classification procedure based solely on $\underline{x}_n$ and the knowledge of the family F of all translates of a given prototype density f, which yields an expected proportion of errors R_n such that

$$R_n - R^*(½, f_1, f_2) \to 0 ,$$

for any $f_1, f_2 \in F$. Thus the underlying densities f_1 and f_2 are effectively learned from $\underline{x}$.

Fralick [17] considers a learning system in which the prior probability η is unknown and the f's $\in F$ are known only up to some unknown parameters. The unknown parameters are treated as random variables; that is, it is assumed that the unknown parameters are distributed according to a certain joint prior distribution. The posterior distribution function is updated after each observation x_i by use of Bayes' rule. The formulation is entirely Bayesian. Procedures are exhibited, based on knowledge of $\underline{x}_n$, achieving $R_n - R^*(\eta, f_1, f_2) \to 0$. The novelty in Fralick's approach consists of his formulation of a sequential decision procedure which requires minimal memory of past observations at each stage. The work in Section II was motivated by similar memory-reducing considerations. A summary of Bayesian learning may be found in Spragins [8].

Hancock and Patrick [18] consider an unsupervised identification problem in which little or nothing is known about the family of probability density functions F. It is assumed that $\underline{x}_n$ but not $\underline{\theta}_n$ is known. The sample space X is divided into a finite number of bins and the resulting empirical multinomial distribution is calculated. From this, estimates $\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_k, \hat{f}_1, \hat{f}_2, \ldots, \hat{f}_k$ are formed of the mixing priors and densities. A Bayes classification rule is used with respect to these estimates.

There are two weaknesses in this procedure. First, no allowance is made for the number of bins to grow with the number of samples. Consequently R_n does not in general converge to the Bayes risk $R^*(\underline{\eta}, \underline{f})$. Second, multinomial distributions are inherently unidentifiable (even when F is identifiable), and it is necessary for the authors to demand that $2k-1$ samples be drawn at a time from the same unknown distribution. These objections may be overcome by further development of this approach in terms of the work of Parzen [19] and Cacoullos [20], with regard to the question of bin growth; and Yakowitz and Spragins [21] and Cooper [22], with regard to the identifiability problems for F.

IV.4 COMPOUND SEQUENTIAL BAYES LEARNING

The learning without a teacher problem may be compared with parallel studies in statistics by Robbins [23], Samuel [24,25], Van Ryzin [26], Swain [27], and Johns [28]. In most of this work it is *not* assumed that $\theta_1, \theta_2, \ldots, \theta_n$ are independent identically distributed random variables. In fact, there is no probability distribution assumed on the θ_i sequence, in any sense. Nevertheless, sequential procedures will be described which asymptotically achieve the Bayes risk defined on the *empirical* distribution of the actual θ_i sequence encountered. These procedures are commonly referred to as compound sequential Bayes rules following the initial work of Robbins and Hannan [29]. However, since the traditional Bayes structure is lacking, namely with regard to the assumption of a prior distribution on the θ_i's, the problem is essentially non-Bayesian, or, at best, empirical Bayesian. (We remark that the designation "empirical Bayes" refers to a slightly different problem, also introduced and studied by Robbins [Refs. 30–44].)

Samuel [24] considers the compound sequential decision problem in which, at the time of the k^th^ decision, the statistician has knowledge only of $\underline{x}_k = (x_1, x_2, \ldots, x_k)$ and $\underline{f} = (f_1, f_2)$. No distribution is assumed on the θ_i sequence. She exhibits a sequential decision rule $\delta_k(\underline{x}_k, \underline{f})$ which achieves a probability of error $R_n = E\{(1/n)\, \Sigma_1^n L_i\}$ such that $R_n - R^*(\overline{\theta}_n, f_1, f_2) \to 0$ as $n \to \infty$, uniformly in all θ sequences in $\{0,1\}^\infty$, where $\overline{\theta}_n$ equals the proportion of 1's in $\underline{\theta}_n$. The decision rule used is essentially the Bayes decision rule with respect to a consistent estimate of the empirical prior probability $\overline{\theta}_n$ given by Hannan and Robbins [29]. When $R^*(\eta, f_1, f_2)$ is not a differentiable function of η, the decision rule involves some artificial randomization. When R^* is differentiable, the observations x_i effectively provide the needed randomization, and a deterministic decision rule may be used.

Obviously, in the Bayesian formulation with θ_i's i.i.d., $\Pr\{\theta_i = 1\} = \eta$, it follows that $\overline{\theta}_n \to \eta$ with probability one. Consequently $R^*(\overline{\theta}_n, f_1, f_2) \to R^*(\eta, f_1, f_2)$ with probability one, and R_n approaches the true Bayes risk $R^*(\eta, f_1, f_2)$. Thus Samuels' work demonstrates that, with a small amount of care, possibly involving some artificial randomization, a Bayes-like decision rule will perform optimally in a much wider class of problems than had been previously thought possible [38]. This is characteristic of the compound Bayes approach. In a subsequent paper [25] Samuel obtains stronger results and demonstrates convergence with probability one of $(1/n)\, \Sigma_1^n L_i - R(\overline{\theta}_n, \underline{f})$ to zero. Robbins [23] extends the results of Samuel to a finite number of states of nature (range of θ) in a general discussion of the empirical Bayes approach to statistics. Tainiter [45] shows that, for an appropriate family of decision rules, $R_n - R^*(\hat{p}_n(\theta_j | \theta_{j-1}, \theta_{j-2}, \ldots, \theta_{j-k}), f_1, f_2)$ converges to zero, for any k, where $\hat{p}_n$ is the *empirical* k^th^ order

Markov dependence exhibited by the sequence $\underline{\theta}_{n-1}$. Swain [27] investigates an estimation problem of similar structure. Thus Markov dependence, real or imagined, may be learned.

Van Ryzin [26] considers the compound and sequential compound decision problem with a finite number of states of nature and actions. Van Ryzin makes precise the artificial randomization and differentiability requirements in Samuel [24] and shows that the rate of convergence of R_n is $0(n^{-½})$.

Van Ryzin [46, 47, 48] is able to dispense with knowledge of the underlying densities $\underline{f}$ at the cost of introducing knowledge of $\underline{\theta}_n$. This is the so-called learning with a teacher problem in which the underlying distributions are unknown. Using Parzen [19] probability density function estimators, Van Ryzin finds decision rules $\delta_k(\underline{x}_k, \underline{\theta}_{k-1})$ for which $R_n - R^*(\overline{\theta}_n, \underline{f}) \to 0$, for all θ sequences in $\{0,1\}^\infty$. Johns [28] considers the two-action compound decision problem extended to an infinite number of states of nature and establishes rates of convergence of R_n to the Bayes risk defined on the m^{th} order empirical joint distribution on the states of nature.

The nonparametric unsupervised learning problem in the compound Bayes framework is considered by Alens and Cover [49,50]. Let $\mathcal{F}$ be a family of probability densities, perhaps infinite in number. Nature is allowed to choose an arbitrary set of k probability density functions $\underline{f} = (f_1, f_2, \ldots, f_k)$ where $f_i \epsilon \mathcal{F}$. These are then fixed for all time. She may then choose an arbitrary sequence of categories $\underline{\theta}_n$. The statistician has knowledge only of $\underline{x}_n$ at the time of his decision. After n observations, it is desired to partition $\{x_1, x_2, \ldots, x_n\}$ into k sets (or categories) with minimum probability of misclassification with respect to the true partition induced by $\{\theta_1, \theta_2, \ldots, \theta_n\}$.

For the sake of comparison, consider the statistician who has complete knowledge, not only of $\underline{x}_n$, but of $\underline{f}$ and $\overline{\underline{\theta}}_n$, where $\overline{\underline{\theta}}_n$ denotes the empirical frequencies of occurrence of the categories in $\underline{\theta}_n$. His minimum expected proportion of misclassifications, over all partitioning procedures on $\underline{x}_n$, will be denoted by $R^*(\overline{\underline{\theta}}_n, \underline{f})$. This is the Bayes probability of error, in the one-shot case, with respect to densities $\underline{f}$ and prior $\underline{\theta}_n$.

Under certain identifiability and completeness conditions [22] on $\mathcal{F}$ a decision rule is exhibited by Alens and Cover based only on $\underline{x}_n$, which incurs a risk R_n for which $R_n - R^*(\overline{\underline{\theta}}_n, \underline{f}) \to 0$, for every $\underline{f} \epsilon \underline{F}$ and every infinite sequence $\theta_1, \theta_2, \ldots$. In other words, the penalty the statistician pays for having no knowledge of the classifications or of the underlying statistics is asymptotically negligible, even in the worst possible cases. This represents a generalization of the work of Fralick, Hancock and Patrick, and Cooper and Cooper in that no probability distribution is assumed on the category sequence $\underline{\theta}_n$. Thus this rule is asymptotically Bayes even in non-Bayesian problems. The compound Bayes work of Van Ryzin is also generalized, from cases in which $\underline{x}_n$ is known and either $\underline{f}$ or $\overline{\underline{\theta}}_n$ is known, to the case in which neither $\underline{f}$ nor $\overline{\underline{\theta}}_n$ is known.

Shubert [51], [52] has examined a problem of the compound Bayes type and obtained very exciting results. Consider a decision-maker who knows only the space D of his possible decisions and an observation space $\mathcal{X}$. In particular, he does not know $\mathcal{F}$ nor does he know the range of θ (which may be infinite). Upon the observation of the i^{th} sample x_i, a decision $d \epsilon D$ is made and a loss $L(\theta_i, d_i)$ is incurred. The loss function itself may, moreover, be random with an unknown distribution depending, of course, on θ and d. No statistics whatsoever are assumed on the sequence of θ_i's. After each decision d_i is made, the decision-maker is told the value of the random loss incurred by him.

Under mild assumptions Shubert [52] is able to demonstrate a sequential decision procedure based on $\underline{x}_n$ and $\underline{L}_n$ which has an average risk which approaches the Bayes risk R^* of the underlying problem evaluated at the hypothetical prior distribution on θ equal to the empirical distribution of the sequence of parameters $\underline{\theta}_n$. Thus, the decision-maker is performing as well asymptotically as if he knew *i)* the distributions $f_\theta(x)$; *ii)* the random loss structure $L(\theta,d)$; and the (possibly time-varying) empirical "statistics" of the parameter sequence $\theta_1, \theta_2, \ldots$.

V. Concluding Remarks

The mathematical theory of pattern recognition, I feel, is differentiated from the classical theory of statistics in that pattern recognition problems do not generally lend themselves to precise statistical formulations. In this sense the field of nonparametric statistics lies closest in spirit to practical problems in pattern recognition. One role that statistics rightfully plays in pattern recognition is that of testing various ad hoc procedures under carefully specified conditions, thus answering the question, "When is this procedure good?" or "What is the natural class of problems for which this procedure is useful?"

For this reason, we have emphasized throughout this paper various "nonstatistical" procedures. By this we mean procedures which are insensitive to the underlying statistical structure, if any, and are robust in the sense that moderate deviations from the underlying assumptions do not invalidate the qualitative goodness of the procedure. Thus the nearest-neighbor classification ideas mentioned in section III are nonstatistical in the sense that the description of the procedure is solely in terms of the data and depends in no way on the underlying distributions. Nevertheless, when there is an underlying statistical structure, it has been demonstrated that procedures of the nearest-neighbor type are good. Since this goodness is insensitive to the underlying statistics, the nearest-neighbor procedure may also be considered robust in the sense just mentioned.

Similarly, the compound Bayes procedures discussed in Chapter IV are nonstatistical, to a lesser extent, in the sense that the crucial assumption of an underlying distribution on the sequence of nature's choices of classifications is shown to be unnecessary for the compound Bayes procedure to have good behavior with respect to a similar procedure in a closely related problem in which that assumption is made and is utilized.

Finally, we have made some remarks in Chapter III about the problem of letting the data choose the best procedure for use on future data.

Although our remarks on this problem have been quite elementary, we hope that there will be much more of a theoretical nature which can be said. In any case, we hope that experimentalists in pattern recognition will take a careful analytical point of view about their reasons for recommending one pattern recognition scheme over another.

REFERENCES

[1] Cover, Thomas M., "Hypothesis Testing with Finite Statistics", to appear, *Ann. Math. Stat.*, June, 1969. Based on paper "Learning with Finite Memory" delivered at IEEE International Conference on Information Theory, San Remo, Italy, September 1967.

[2] Cover, Thomas M., and Martin E. Hellman, "A Solution of the Two-Armed Bandit with Finite Memory," to be submitted to *IEEE Transactions on Inf. Theory, 1968.*

[3] Robbins, H., "A Sequential Decision Problem with Finite Memory", *Proc. Nat. Acad. Sci., 42,* 920–923, 1956.

[4] Smith, C. V. and Pyke, R., "The Robbins-Isbell Two-Armed Bandit Problem with Finite Memory", *Annals of Mathematical Statistics, 36,* 1375–1386, 1965.

[5] Cover, Thomas M., "A Note on the Two-Armed Bandit Problem with Finite Memory", IBM Yorktown Heights Technical Report RC-1913, October 1967, *Information and Control,* May-June, 1968.

[6] Dynkin, E.B., "Necessary and Sufficient Conditions for a Family of Probability Distributions", in *Selected Translations in Mathematical Statistics and Probability, 1,* 17–40, 1961.

[7] Abramson, N. and Braverman, D., "Learning to Recognize Patterns in a Random Environment", *IEEE Transactions on Information Theory, IT–8,* 58–63, 1962.

[8] Spragins, J., "Learning Without a Teacher", *IEEE Transactions on Information Theory, IT–12,* 223–230, 1966.

[9] Denny, J.L., "On Continuous Sufficient Statistics", *Annals of Mathematical Statistics.*

[10] Denny, J.L., "A Continuous Real-Valued Function on E^n Almost Everywhere 1-1", *Fund. Math.*

[11] Cover, Thomas M., and Hart, Peter, "Nearest Neighbor Pattern Classification", *IEEE Trans. on Information Theory, IT–13,* 21–27, 1967.

[12] Cover, Thomas M., "Estimation by the Nearest Neighbor Rule", *IEEE Trans. on Information Theory,* January 1968.

[13] Cover, Thomas M., "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications to Pattern Recognition", *IEEE Trans. on Electron. Computers, EC–14,* June 1965.

[14] Cover, Thomas M., "Rates of Convergence of Nearest Neighbor Decision Procedures" paper delivered at First Annual Hawaii International Conference on Systems Theory, January 1968.

[15] Whitney, A.W., Ph.D. Dissertation, University of Missouri, Columbia, Missouri, January 1967.

[16] Cooper, D. B. and Cooper, P.W., "Nonsupervised Adaptive Signal Detection and Pattern Recognition", *Inform. Control, 7,* September 1964.

[17] Fralick, S.C., "Learning to Recognize Patterns Without a Teacher", Rept. SEL-65-011 (TR No. 6103-10), Stanford Electronics Laboratories, Stanford, California, March 1965.

[18] Hancock, J.C. and Patrick, E.A., "Learning Probability Spaces for Classification and Recognition of Patterns With or Without Supervision", TR No. EE 65-21, School of Electrical Engineering, Purdue University, Lafayette, Indiana, September 1965.

[19] Parzen, E., "On Estimation of a Probability Density Function and Mode", *Ann. Math. Stat., 33,* September 1962.

[20] Cacoullos, T., "Estimation of a Multivariate Density", TR No. 40, Department of Statistics, University of Minnesota.

[21] Yakowitz, S. and Spragins, J., "On the Identifiability of Finite Mixtures", *Ann. Math. Stat., 39,* February 1968.

[22] Cooper, D.B., "On the Existence of Nonsupervised Adaptive Signal Detectors; and Detector Estimation Using Stochastic Approximation Methods", Ph.D. Thesis, Columbia University, New York, April 1966.

[23] Robbins, H., "The Empirical Bayes Approach to Statistical Decision Problems", *Ann. Math. Stat., 35,* February 1964.

[24] Samuel, E., "Asymptotic Solutions of the Sequential Compound Decision Problem" *Ann. Math. Stat., 34,* September 1963.

[25] Samuel, E., "Convergence of the Losses of Certain Decision Rules for the Sequential Compound Decision Problem", *Ann. Math. Stat., 35,* December 1966.

[26] Van Ryzin, J.R., "The Compound Decision Problem with m x n Finite Loss Matrix", *Ann. Math. Stat.,* 37, April 1966.

[27] Swain, D.D., "Bounds and Rates of Convergence for the Extended Compound Estimation Problem in the Sequence Case", TR No. 81, Department of Statistics, Stanford University, Stanford, California, June 1965.

[28] Johns, M.V. Jr., "Two-Action Compound Decision Problems", TR No. 87, Department of Statistics, Stanford University, Stanford, California, March 1966.

[29] Hannan, J.F. and Robbins, H., "Asymptotic Solutions of the Compound Decision Problem for Two Completely Specified Distributions", *Ann. Math. Stat., 26,* January 1955.

[30] Choi, K., "A Note on the Empirical Bayes Approach to Statistical Decision Problems", Mathematical Sciences Technical Report No. 19, University of Missouri, 1966.

[31] Choi, K., "Composite vs. Composite Hypothesis Testing (Empirical Bayes Approach)", Mathematical Sciences Technical Report No. 20, University of Missouri, 1967.

[32] Deely, J.J., "Multiple Decision Procedures from an Empirical Bayes Approach", Mimeograph Series No. 45, Purdue University, Department of Statistics, 1965.

[33] Johns, M.V. Jr., "Contributions to the Theory of Non-Parametric Empirical Bayes Procedures in Statistics", Ph.D. Dissertation, Columbia University, 1956.

[34] Johns, M.V. Jr., "Non-Parametric Empirical Bayes Procedures", *Ann. Math. Statistics, 28,* 649–669, 1957.

[35] Johns, M.V. Jr., "An Empirical Bayes Approach to Non-Parametric Two-Way Classification", *Studies in Item Analysis and Prediction,* Stanford University, Stanford, California, 1961.

[36] Kagan, A.M., "An Empirical Bayes Approach to the Estimation Problem" *Doklady of the Academy of Sciences of the USSR, 147,* 1020–1021, 1962.

[37] Miyasawa, K., "An Empirical Bayes Estimator of the Mean of a Normal Population", *Bull. International Statist. Inst. 38*, 181–188, 1961.

[38] Neyman, J., "Two Breakthroughs in the Theory of Statistical Decision Making", *Review International Statist. Inst. 30*, 11–27, 1962.

[39] Robbins, H., "An Empirical Bayes Approach to Statistics", *Proceedings of Third Berkeley Symposium on Statistics and Probability*, 157–164, 1955.

[40] Robbins, H., "The Empirical Bayes Approach to Testing Statistical Hypotheses", *Review of International Statist. Inst., 31*, 195–208, 1963.

[41] Rutherford, J.R., "The Empirical Bayes Approach: Estimation of Posterior Quantiles", Unpublished Paper. Royal Military College, Kingston, Ontario, Canada, 1966.

[42] Rutherford, J.R. and Krutchkoff, R.G., ' The Empirical Bayes Approach: Estimating the Prior Distribution", Unpublished Paper. V.P.I., 1966.

[43] Samuel, E., "An Empirical Bayes Approach to the Testing of Certain Parametric Hypotheses", *Ann. Math. Stat., 34*, 1370–1385, 1963.

[44] Sastry, R.M.V., "Empirical Bayes Estimation in Regression Analysis", Calif. State College at Fullerton, Unpublished Paper, 1966.

[45] Tainiter, M., "Sequential Hypothesis Tests for the R-dependent Marginally Stationary Processes", *Ann. Math. Stat., 37*, February 1966.

[46] Van Ryzin, J.R., "Non-Parametric Bayesian Decision Procedures for (Pattern) Classification with Stochastic Learning", *Transactions of the Fourth Prague Conf. on Information Theory, Statistical Decision Functions, Random Processes*, Prague, Czechoslovakia, 1965.

[47] Van Ryzin, J.R., "Repetitive Play in Finite Statistical Games with Unknown Distribution", *Ann. Math. Statist., 37*, 976–994, 1966.

[48] Van Ryzin, J.R., "Bayes Risk Consistency of Classification Procedures Using Density Estimation", *Sankhyā, Series A, 28*, 1966.

[49] Alens, N., "Compound Bayes Learning Without a Teacher", Rept. SU-SEL-67-019 (TR No. 6151-2), Stanford Electronics Laboratories, Stanford, California, August 1967.

[50] Alens, N. and Cover, Thomas M., "Compound Bayes Learning Without a Teacher" (Abstract) Proceedings of the First Annual Princeton Conference on Systems Theory, April 1967.

[51] Shubert, B.O., "Repetitive Play of an Unknown Game Against Nature", SEL Tech. Rept. No. 6151-3, Stanford University, Stanford, California, December 1967.

[52] Shubert, B.O., "Learning With a Lack of Prior Data", SEL Tech. Rept. No. 6151-4, Stanford University, Stanford, California, December 1967, submitted to *Ann. Math. Stat.*

PARALLEL COMPUTATION IN PATTERN RECOGNITION

Michael J. B. Duff

UNIVERSITY COLLEGE LONDON
LONDON, ENGLAND

1. Introduction

One of the rather curious factors hindering any coordinated attack on the problems of pattern recognition is the difficulty experienced in reaching agreement as to what is meant by a 'pattern'. Because the word has a common everyday usage, its significance is only loosely defined and tends to be appreciated more intuitively than analytically.

The need to give some precision to the definition of the word is felt as soon as one attempts to devise a pattern recognition system. Sooner or later, the question is invariably asked: 'Does the system have any generality or will it only recognise patterns in a narrowly defined field?'

It is convenient, if not essential, to regard any pattern recognition system as including in its early stages some form of input device which will represent the input as an ordered set of numbers. The input may be a train of sound waves, a projected optical image of a page of writing, or a view out of a satellite window; it can be assumed that the essential quality of 'pattern' in the input can be translated into the ordered number set by means of microphones or photocells, together with amplifiers, digitizing electronics, or whatever apparatus may prove necessary. The actual method used, although it may require the application of very sophisticated techniques and a depth of understanding of the properties of the input which may not be readily available, is irrelevant to the present argument. The important assumption is that the input device will not destroy any of that part of the input information which is necessary for the pattern to be recognised by analysis of the ordered number set. Accepting the validity of this assumption, we may now refer to the ordered number set itself as 'the pattern', allowing us to ignore physical peculiarities of the original input. In order to fix our ideas, let us assume that the numbers are arranged in a square array with twenty rows and twenty columns. Initially, and for the sake of simplicity, we may further assume that the numbers can take only the values one or zero.