The Annals of Statistics
1977, Vol. 5, No. 1, 185–190

CLASSIFICATION BY MAXIMUM POSTERIOR PROBABILITY

BY C. P. SHAPIRO[1]

Michigan State University

The problem of classifying *each* of n observations to one of two subpopulations is considered. The classification rule examined chooses that classification with maximum posterior probability. Limiting behavior of the rule is given and several examples are presented which show that the rule can lead to classifying all observations to the same subpopulation. Three simulation studies are reported to indicate that this extreme behavior may occur in small samples.

1. Introduction of the problem. We take a Bayesian approach to the problem of classifying *each* of n observations to one of two subpopulations. Let θ be a random variable in $[0, 1]$ with distribution function Λ. Let $\mathbf{Z} = (Z_1, \cdots, Z_n)$ be a random vector with $P(\mathbf{Z} = \mathbf{z} \mid \theta) = \prod_{i=1}^n \theta^{z_i}(1 - \theta)^{1-z_i}$. Let G and H be known univariate distribution functions and suppose random vector $\mathbf{X} = (X_1, \cdots, X_n)$ can be observed where

$$P(X_1 \leqq x_1, \cdots, X_n \leqq x_n \mid \theta, \mathbf{Z} = \mathbf{z}) = \prod_{i=1}^n G(x_i)^{z_i} H(x_i)^{1-z_i} .$$

The problem is to classify each observed X_i to G or to H, where the $\mathbf{Z}$ above is the true classification of $\mathbf{X}$. This problem differs from the usual 2-choice multivariate classification problem in that X_i and X_j can come from different distributions if $i \neq j$. The problem is a special case of the n-variate classification problem with 2^n classes. The criterion for classification will be maximum posterior probability.

Without loss of generality, assume G and H have densities g and h with respect to a sigma finite measure. Then using Bayes' theorem, the posterior probability of the classification $\mathbf{z}$ is

$$P(\mathbf{Z} = \mathbf{z} \mid \mathbf{X} = \mathbf{x}) = cE(\theta^{\Sigma z_i}(1 - \theta)^{n - \Sigma z_i}) \prod_{i=1}^n g(x_i)^{z_i} h(x_i)^{1-z_i}$$

where c is the normalizing constant. Let $\mathbf{a} = (a_1, \cdots, a_n)$ denote the action vector where $a_i = 1$ indicates classifying x_i to g and $a_i = 0$ indicates classifying x_i to h. The maximum posterior probability (m.p.p.) rule chooses the action $\mathbf{a}$ which maximizes $P(\mathbf{Z} = \mathbf{a} \mid \mathbf{X} = \mathbf{x})$.

Throughout the paper assume that g and h are mutually absolutely continuous so that the m.p.p. rule chooses the action $\mathbf{a}$ which maximizes

$$[\prod_{i=1}^n (g(x_i)/h(x_i))^{a_i}] E\theta^{\Sigma a_i}(1 - \theta)^{n - \Sigma a_i} .$$

Several properties of the rule above become obvious with the introduction of

Received October 1974; revised May 1976.

[1] Parts of the paper are from the author's dissertation at the University of Michigan.

AMS 1970 *subject classifications*. Primary 62C10; Secondary 62E20.
Key words and phrases. Bayesian, classification.

new notation. Let $W_1 \leqq \cdots \leqq W_n$ be the ordered values of the likelihood ratios $g(X_i)/h(X_i)$, and let $Z_i^* = Z_j$ and $a_i^* = a_j$ when $g(X_j)/h(X_j) = W_i$. Then the m.p.p. rule chooses the action $\mathbf{a}^*$ which maximizes

$$(\prod W_i^{a_i^*}) E\{\theta^{\Sigma a_i^*}(1-\theta)^{n-\Sigma a_i^*}\}.$$

The action $\mathbf{a}^*$ with maximum posterior probability always has the form $(0, \cdots, 0, 1, \cdots, 1)$ since W_i increases in i. Thus, the rule makes one cut in the ordered values of the likelihood ratios, $g(x)/h(x)$, and classes to g any X_j yielding a ratio value above that cut. A rule with this property is completely determined by the number of observations classed to g. If k_n observations are classed to g, these observations must give the top k_n values of the likelihood ratio. The limiting properties of such a rule can also be studied in terms of k_n.

2. Limit results. For t in $[0, 1]$, define

$$\phi_n(t) = \phi_{1n}(t) + \phi_{2n}(t)$$

where

$$\begin{aligned}\phi_{1n}(t) &= n^{-1} \textstyle\sum_{i=n-[nt]+1}^{n} \log W_i\,, \quad && t \geqq 1/n \\ &= 0\,, && t < 1/n\,, \\ \phi_{2n}(t) &= n^{-1} \log E\{\theta^{[nt]}(1-\theta)^{n-[nt]}\}\end{aligned}$$

and $[\cdot]$ is the greatest integer function. The m.p.p. rule classifies k_n observations to g (corresponding to the top values of the likelihood ratios) where $k_n = nt_n$ and $t_n = \inf\{t\colon \phi_n(t) \geqq \phi_n(t') \text{ for all } t'\}$. Note that t_n is simply the proportion of observations classed to g, and k_n is the number of observations classified to g (k_n is an integer). The limiting form of the rule is given by the limit of t_n which will be derived from the limit of $\phi_n(t)$.

Define $\phi(t, \theta) = \phi_1(t, \theta) + \phi_2(t)$ where

$$\begin{aligned}\phi_1(t, \theta) &= E_\theta[\log Y \| \{Y \geqq \tilde{F}_\theta^{-1}(1-t)\}] \\ \phi_2(t) &= \log \|\theta^t(1-\theta)^{1-t}\|_\infty\,.\end{aligned}$$

Above, $\tilde{F}_\theta$ is the distribution of $Y = g(X)/h(X)$ when X is distributed F_θ, $\|\{\cdot\}$ is the set indicator function, $E_\theta[\cdot]$ is expectation under distribution $\tilde{F}_\theta$, and $\|\cdot\|_\infty$ is the essential sup norm.

The main limit theorem is given below.

THEOREM 1. *Fix θ in $[0, 1]$ and assume the prior distribution Λ puts zero probability on the points 0 and 1. Assume that $\tilde{F}_\theta$ is continuous and strictly increasing with $\int |\log y|\, d\tilde{F}_\theta(y) < \infty$. Then, given θ,*

(i) $\phi_{1n}(t) \to \phi_1(t, \theta)$ a.s. *and uniformly in t.*

(ii) $\phi_{2n}(t) \to \phi_2(t)$ *uniformly in t, and hence* $\phi_n(t) \to \phi(t, \theta)$ a.s. *and uniformly in t.*

(iii) *Furthermore, if for fixed θ, $\phi(t, \theta)$ assumes a unique maximum at $t_0 = t_0(\theta)$, then $t_n \to t_0$ a.s. given θ.*

Part (iii) of the theorem follows in the usual way from parts (i) and (ii) and the assumption of a unique maximum. Parts (i) and (ii) require some work and each part will be considered separately.

For part (i), Lemmas 1 and 2 are needed. Lemma 1 is a standard result and will be stated without proof.

LEMMA 1. *Let* $Y_1, \cdots, Y_n$ *be* i.i.d. *with continuous distribution* F. *Let* $W_1, \cdots, W_n$ *be the corresponding order statistics. Let* $Q(y)$ *be a nonnegative measurable function of* y *such that* $EQ(Y) < \infty$. *Define* $\sum_{i=n+1}^{n} Q(W_i)$ *equal to zero. Then for all* t *in* $[0, 1]$,

$$n^{-1} \sum_{i=n-[nt]+1}^{n} Q(W_i) \to EQ(Y) \| \{A_t\} \quad \text{a.s.},$$

where $A_t = \{Y \geqq F^{-1}(1 - t)\}$.

LEMMA 2. *Let* $X_1, X_2, \cdots$ *be random variables and let* $\mathbf{X}_n = (X_1, \cdots, X_n)$. *Let* $Q_n(\mathbf{x}_n, t)$ *be a measurable function of* $\mathbf{x}_n$ *for each* t *in* $[0, 1]$ *and increasing in* t *for each* $\mathbf{x}_n$. *Let* $Q(t)$ *be a continuous function of* t. *Then*

$$Q_n(\mathbf{X}_n, t) \to Q(t) \quad \text{a.s. } \textit{for each } t \textit{ in } [0, 1]$$

implies

$$\sup_{0 \leqq t \leqq 1} |Q_n(\mathbf{X}_n, t) - Q(t)| \to 0 \quad \text{a.s.}$$

PROOF. The proof is the same as that for nonrandom Q_n. See Breiman (1968).

PROOF OF THEOREM 1, PART (i). Recall that $\tilde{F}_\theta$ is the distribution of Y.

Fix θ in $[0, 1]$ and let $A_t = \{Y \geqq \tilde{F}_\theta^{-1}(1 - t)\}$. Given θ, the random variables $Y_1, \cdots, Y_n$ are i.i.d. $\tilde{F}_\theta$ and $W_1, \cdots, W_n$ are the corresponding order statistics. Let $(\cdot)^+ = \max\{(\cdot), 0\}$. Lemma 1 implies that for every t in $[0, 1]$,

$$n^{-1} \sum_{i=n-[nt]+1}^{n} (\log W_i)^+ \to E_\theta(\log Y)^+ \| \{A_t\} \quad \text{a.s. given } \theta .$$

This limit function is continuous in t by the assumptions on $\tilde{F}_\theta$ and the function is strictly increasing. Also, the function involving the summation is increasing in t. Thus, the convergence is uniform in t by Lemma 2. Applying the same argument to

$$n^{-1} \sum_{i=n-[nt]+1}^{n} (\log W_i)^- ,$$

where $(\cdot)^- = \max\{-(\cdot), 0\}$, concludes the proof of part (i).

For part (ii), note that $\phi_{2n}(t)$ is not random and write

$$\phi_{2n}(t) = \log \left(E\{\theta^{[nt]}(1 - \theta)^{n-[nt]}\}\right)^{1/n} .$$

Ignoring the $[\cdot]$ function, the above is equal to $\log \|q_t\|_n$ where $q_t(\theta) = \theta^t(1 - \theta)^{1-t}$ with $q_0(0) = 1 = q_1(0)$ and where $\|\cdot\|_n$ is the n-norm with respect to the prior measure on $[0, 1]$. The convergence of $\|q_t\|_n$ will be studied first. By properties of norms, $\|q_t\|_n \to \|q_t\|_\infty$ as $n \to \infty$, and thus, uniform convergence in t is the main problem.

LEMMA 3. *If the prior distribution puts zero probability on the points* 0 *and* 1, *then for* t *in* $[0, 1]$

(i) $||q_t||_n$ *is continuous in* t,
(ii) $||q_t||_\infty$ *is continuous in* t.

PROOF. See Shapiro (1972).

The lemma above allows use of Dini's theorem (Lemma 4) to prove uniform convergence of $||q_t||_n$.

LEMMA 4 (Dini's theorem). *If* $\{K_n\}_1^\infty$ *is a sequence of real-valued continuous functions converging pointwise to a continuous limit function* K *on* $[0, 1]$ *and if* $K_n(t) \leqq K_{n+1}(t)$ *for each* t *in* $[0, 1]$ *and all* n, *then* K_n *tends to* K *uniformly on* $[0, 1]$.

PROOF. See Apostol, page 425 (1957).

PROOF OF THEOREM 1, PART (ii). By properties of norms, $||q_t||_n \to ||q_t||_\infty$ and $||q_t||_n \leqq ||q_t||_{n+1}$ for each t in $[0, 1]$. By Lemma 3, all of these functions are continuous in t. It is easy to show that

$$\inf\{||q_t||_n : 0 \leqq t \leqq 1,\ 1 \leqq n < \infty\} > 0$$

when the prior distribution puts zero probability on points 0 and 1. Thus, $\log ||q_t||_n$ is a monotone sequence of continuous functions converging pointwise to a continuous limit, $\log ||q_t||_\infty$, on $[0, 1]$. Dini's theorem thus implies the convergence is uniform in t.

To prove $\phi_{2n}(t)$ converges uniformly on $[0, 1]$, it suffices to show that if $t_n' \to t_0$ then $\phi_2(t_0) = \log ||q_{t_0}||_\infty$. Let $t_n'' = [nt_n']/n$. Then $t_n'' \to t_0$ and $\phi_{2n}(t_n') = \log ||q_{t_n''}||_n$. This last expression tends to $\log ||q_{t_0}||_\infty$ by the uniform convergence of $\log ||q_t||_n$ proven above.

3. Examples. Assume the prior distribution has support $[0, 1]$ and that the conditions in Theorem 1 hold. Then

$$\phi(t, \theta) = \phi(t) = \int_{1-t}^1 \log \tilde{F}_\theta^{-1}(u)\, du + t \log t + (1 - t) \log (1 - t)$$

and

$$\phi'(t) = \log\left(\frac{t}{1 - t}\right) \tilde{F}_\theta^{-1}(1 - t)\,.$$

Hence,

$$\phi'(t) \gtreqless 0 \quad \text{if and only if} \quad \tilde{F}_\theta(1 - t/t) \gtreqless 1 - t\,. \tag{3.1}$$

The classification rule is said to be degenerate in the limit, or degenerate, if t_n, the proportion classified to g, tends to 0 or 1 a.s. for each θ. If density g is assumed decreasing and h is assumed uniform, sufficient conditions for degeneracy and nondegeneracy can be given (Shapiro, 1972). However, most densities of interest are not of this type and hence most problems must be considered individually.

Three examples follow.

EXAMPLE 1. *Nondegenerate rule.* Let $g(x) = a_p[x^{-p}(2 - x)^{-p} - 1], 0 < x < 1$,

and let $h(x) = 1$, $0 < x < 1$. If $\frac{1}{2} < p < 1$, then it can be shown that $\phi'(t) > 0$ in the neighborhood of 0 and < 0 in the neighborhood of one.

EXAMPLE 2. *Degenerate rule.* Let $g(x) = 2(1 - x)$ and $h(x) = 2x$, $0 < x < 1$. Then

$$t_0(\theta) = 0 \quad \text{if} \quad \theta < \tfrac{1}{2},$$
$$= 1 \quad \text{if} \quad \theta > \tfrac{1}{2}, \qquad \text{and}$$

when $\theta = \frac{1}{2}$, $\phi(t, \theta)$ is constant.

EXAMPLE 3. *Mixture of exponentials.* Let $g(x) = \alpha e^{-\alpha x}$ and $h(x) = \beta e^{-\beta x}$, with $0 < \beta < \alpha$, $0 < x$. Let $\gamma = \alpha/\beta$ and express (3.1) as

$$\theta(1 - t) + (1 - \theta)\gamma t \gtreqless (1 - t)^{\gamma - 2/\gamma - 1}(\gamma t)^{\gamma/\gamma - 1}. \tag{3.2}$$

If $\gamma \leqq 2$, 0 and 1 are the only local maxima since $\phi(t)$ assumes a unique minimum in (0, 1) for each θ in (0, 1). Thus, the rule is degenerate.

Suppose $\gamma > 2$. Then examination of (3.2) yields the following conclusion: $\phi(t)$ has two local maxima, one at $t = 0$ and another at $t_0 > \frac{2}{3}$. Such behavior occurs even when the two exponentials are widely separated since $\gamma = \alpha/\beta$ can be as large as desired.

Given below are the results of three simulations for selected α and β values. Samples of size $n = 10$ with $\beta = 1$ and $\theta = \frac{1}{3}$ were generated for values $\alpha = 2$, 5, and 10. The table gives the percentage of samples in which k_n observations were classified to g. When $\alpha = 2$, the m.p.p. rule is degenerate in 98% of the samples. With $\alpha = 5$, the rule is degenerate in 68% of the samples.

Number of samples	k_n											
	α	0	1	2	3	4	5	6	7	8	9	10
100	2	66%	0	0	0	0	0	0	0	0	2%	32%
100	5	62%	0	0	0	2%	6%	9%	6%	7%	4%	4%
165	10	32.9	0	6.1	6.7	11.0	14.0	14.6	7.9	4.3	2.5	0.0%

Thus, the maximum probability rule may behave badly for small samples and may not be a "wise" rule in certain multivariate situations.

4. Acknowledgments. The author wishes to thank Professor Michael Woodroofe for guidance and encouragement during this research and the referee for many helpful comments.

REFERENCES

[1] APOSTOL, T. M. (1957). *Mathematical Analysis.* Addison-Wesley, Reading.
[2] BREIMAN, L. (1968). *Probability.* Addison-Wesley, Reading.
[3] GOFFMAN, C. and PEDRICK, G. (1965). *First Course in Functional Analysis.* Prentice-Hall, Englewood Cliffs.
[4] SHAPIRO, C. P. (1972). Bayesian classification. Unpublished Ph. D. dissertation. Univ. of Michigan.

[5] SYMONS, M. J. (1969). A Bayesian test of normality with a mixture of two normals as the alternative and applications to cluster analysis. Unpublished Ph. D. dissertation. Univ. of Michigan.

DEPARTMENT OF STATISTICS AND PROBABILITY
MICHIGAN STATE UNIVERSITY
WELLS HALL
EAST LANSING, MICHIGAN 48824