ON SOME CLASSIFICATION PROBLEMS—I

By S. JOHN

Indian Statistical Institute, Calcutta

SUMMARY. A classification procedure is described and its performance characteristics are studied. The procedure described is one which can be adopted in situations where the only information on alternative populations is contained in samples.

1. Introduction

In problems of discrimination we have an individual, known to belong to one or other of k populations and we want to assign it to its correct population. Such problems of classification often arise in several branches of science. The technique of discriminant functions which was devised by Fisher (1936) has proved to be invaluable in tackling classification problems. But the construction of the discriminant function is possible only when we know the values of the parameters characterizing the populations to be discriminated between. This raises the question as to what is to be done when such knowledge is absent. In this paper we describe a classification procedure suited to such situations and make a detailed study of the chances of error involved in this procedure.

Let us first consider the case of two populations P_1 and P_2. We shall suppose P_1 and P_2 to be multivariate normal with means $\mu^{(1)} = (\mu_1^{(1)}, \mu_2^{(1)} \ldots, \mu_p^{(1)})$, $\mu^{(2)} = (\mu_1^{(2)}, \mu_2^{(2)}, \ldots, \mu_p^{(2)})$ respectively and a common dispersion matrix Σ. We shall, in this paper, suppose that Σ is known. We have an individual with measurements $x = (x_1, x_2, \ldots, x_p)$ and we want to assign him to P_1 or P_2. To this end we draw samples of size N_1 and N_2 from P_1 and P_2 and perform tests of significance to see whether x could have come from P_1 or P_2. Suppose the test rejects the hypothesis that x came from P_1 at a level α_1 and the hypothesis that x came from P_2 at a level α_2. We assign x to P_1 or P_2 according as $\alpha_1 \lessgtr \alpha_2$. Let $\bar{x}^{(1)}$ be the vector of sample means in the first sample and let $\bar{x}^{(2)}$ be the vector of sample means in the second sample. We may, for the tests of significance, use the criteria $(N_1/N_1+1)(\bar{x}^{(1)}-x)\Sigma^{-1}(\bar{x}^{(1)}-x)'$ and $(N_2/N_2+1)(\bar{x}^{(2)}-x)\Sigma^{-1}(\bar{x}^{(2)}-x)'$ respectively, each being distributed as a chi-square with p degrees of freedom under the respective null hypotheses. In this case our procedure amounts to assigning x to P_1 or P_2 according as

$$(N_1/N_1+1)(\bar{x}^{(1)}-x)\Sigma^{-1}(\bar{x}^{(1)}-x)' \lessgtr (N_2/N_2+1)\,(\bar{x}^{(2)}-x)\,\Sigma^{-1}(\bar{x}^{(2)}-x)' \quad \ldots \quad (1.1)$$

In the following sections we shall study in some detail the performance characteristics of such a procedure. A similar study of other classification procedures will be made in subsequent papers.

2. Probabilities of misclassification

If x really came from P_1 what is the probability of wrongly assigning it to P_2? If x really came from P_2 what is the chance of wrongly assigning it to P_1? In other words, we require

$$Pr(z_1 > z_2 | x \epsilon P_1) \text{ and } Pr(z_1 < z_2 | x \epsilon P_2) \qquad \dots \quad (2.1)$$

where
$$z_1 = (N_1/N_1+1)(\bar{x}^{(1)}-x)\,\Sigma^{-1}(\bar{x}^{(1)}-x)' \qquad \dots \quad (2.2)$$

and
$$z_2 = (N_2/N_2+1)(\bar{x}^{(2)}-x)\,\Sigma^{-1}(\bar{x}^{(2)}-x)'. \qquad \dots \quad (2.3)$$

We shall explain how to evaluate $Pr(z_1 > z_2 | x \epsilon P_1)$; $Pr(z_1 < z_2 | x \epsilon P_2)$ can be evaluated similarly.

Put
$$y^{(1)} = (N_1/N_1+1)^{\frac{1}{2}}(\bar{x}^{(1)}-x)A \qquad \dots \quad (2.4)$$

and
$$y^{(2)} = (N_2/N_2+1)^{\frac{1}{2}}(\bar{x}^{(2)}-x)A \qquad \dots \quad (2.5)$$

where A is a $(p \times p)$ matrix such that $A'\,\Sigma\,A = I$.

We can now write

$$z_1 = y^{(1)}y^{(1)'}, \quad z_2 = y^{(2)}y^{(2)'}. \qquad \dots \quad (2.6)$$

The joint distribution of $(y^{(1)}, y^{(2)})$ is multivariate normal with means $(0, \theta)$ and dispersion matrix

$$\begin{pmatrix} I & \rho I \\ \rho I & I \end{pmatrix}$$

where
$$\theta = \left(\frac{N_2}{N_2+1}\right)^{\frac{1}{2}}(\mu^{(2)}-\mu^{(1)})A \qquad \dots \quad (2.7)$$

and
$$\rho = \frac{N_1^{1/2}N_2^{1/2}}{(N_1+1)^{\frac{1}{2}}(N_2+1)^{\frac{1}{2}}}. \qquad \dots \quad (2.8)$$

We now make one more transformation and put

$$w^{(1)} = y^{(1)}C, \quad w^{(2)} = y^{(2)}C \qquad \dots \quad (2.9)$$

where C is an orthogonal matrix whose first column is

$$\left(\frac{\theta_1}{\left[\sum\limits_{i=1}^{p} \theta_i^2\right]^{\frac{1}{2}}}, \frac{\theta_2}{\left[\sum\limits_{i=1}^{p} \theta_i^2\right]^{\frac{1}{2}}}, \dots, \frac{\theta_p}{\left[\sum\limits_{i=1}^{p} \theta_i^2\right]^{\frac{1}{2}}} \right)$$

In terms of the variables $w^{(1)}$ and $w^{(2)}$,

$$z_1 = w^{(1)}w^{(1)'}, \quad z_2 = w^{(2)}w^{(2)'}. \qquad \dots \quad (2.10)$$

The joint density function of $(w^{(1)}, w^{(2)})$ splits up into p factors and is, in fact,

$$w_1^{(1)}, w_1^{(2)}) \prod_{i=2}^{p} g(w_i^{(1)}, w_i^{(2)}) \qquad \ldots \quad (2.11)$$

where
$$f(a, b) = \frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}}\, e^{-\frac{1}{2(1-\rho^2)}[a^2-2\rho a\,(b-\delta)+(b-\delta)^2]} \qquad \ldots \quad (2.12)$$

$$g(a, b) = \frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}}\, e^{-\frac{1}{2(1-\rho^2)}[a^2-2\rho ab+b^2]} \qquad \ldots \quad (2.13)$$

and
$$\delta = (\sum_{i=1}^{p} \theta_i^2)^{\frac{1}{2}} = \{(N_2/N_2+1)(\mu^{(2)}-\mu^{(1)})\Sigma^{-1}(\mu^{(2)}-\mu^{(1)})'\}^{\frac{1}{2}} \quad \ldots \quad (2.14)$$

δ is thus a simple multiple of the distance between P_1 and P_2.

Now put

$$u = [2(1-\rho)]^{-\frac{1}{2}}(w^{(2)}-w^{(1)}) \qquad \ldots \quad (2.15)$$

$$v = [2(1+\rho)]^{-\frac{1}{2}}(w^{(2)}+w^{(1)}). \qquad \ldots \quad (2.16)$$

Clearly $z_1 > z_2$ if and only if $uv' < 0$. Thus we get the result

$$Pr(z_1 > z_2 | x \in P_1) = Pr(uv' < 0). \qquad \ldots \quad (2.17)$$

We now address ourselves to the task of evaluating $Pr(uv' < 0)$. We observe that $u_1, u_2, \ldots, u_p$; $v_1, v_2, \ldots, v_p$ are all independently distributed normal variables. The variance of each is unity. All of them except u_1 and v_1, have got zero expectation. But

$$E(u_1) = [2(1-\rho)]^{-\frac{1}{2}}\,\delta = \eta \text{ (say)} \qquad \ldots \quad (2.18)$$

and
$$E(v_1) = [2(1+\rho)]^{-\frac{1}{2}}\,\delta = \zeta \text{ (say).} \qquad \ldots \quad (2.19)$$

First we shall find $Pr(uv' < 0|u)$ and finally get $Pr(uv' < 0)$ by finding $E[Pr(uv' < 0|u)]$.

$$Pr(uv' < 0|u) = \int_{-\infty}^{-\zeta u_1(\Sigma u_i^2)^{-\frac{1}{2}}} \frac{1}{(2\pi)^{\frac{1}{2}}}\, e^{-t^2/2}\, dt$$

$$= \int_{\zeta u_1(\Sigma u_i^2)^{-\frac{1}{2}}}^{\infty} \frac{1}{(2\pi)^{\frac{1}{2}}}\, e^{-t^2/2}\, dt. \qquad \ldots \quad (2.20)$$

We now make the substitution

$$u_1 = r\cos t, \quad \chi = (u_2^2+u_3^2+\ldots+u_p^2)^{\frac{1}{2}} = r\sin t. \qquad \ldots \quad (2.21)$$

Then
$$Pr(uv' < 0|u) = \int_{\zeta\cos t}^{\infty} \frac{1}{(2\pi)^{\frac{1}{2}}} \; e^{-t^2/2}\,dt. \qquad \ldots \quad (2.22)$$

We observe that $Pr(uv' < 0|u)$ depends on u only through t. Therefore if $\varphi(t)$ is the density function of t,

$$Pr(uv' < 0) = \int \Big(\int_{\zeta\cos t}^{\infty} (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2} \Big)\, \varphi(t)\,dt. \qquad \ldots \quad (2.23)$$

We have thus to find the distribution of t.

3. Distribution of t

The joint density of u_1 and χ^2 is easily seen to be

$$2^{-(p-1)/2}\Big[\,\Gamma\Big(\frac{p-1}{2}\Big)\Big]^{-1}(\chi^2)^{(p-3)/2}\; e^{-\chi^2/2}(2\pi)^{-\frac{1}{2}}\;\; e^{-\frac{1}{2}(u_1-\eta)^2} \qquad \ldots \quad (3.1)$$

Making the substitution

$$u_1 = r\cos t, \quad \chi = r\sin t \qquad \ldots \quad (3.2)$$

we find that the joint density of r and t is

$$\pi^{-\frac{1}{2}}\, 2^{-(p-2)/2}\Big[\Gamma\Big(\frac{p-1}{2}\Big)\Big]^{-1}\;\; e^{-\eta^2/2}\;\; e^{-r^2/2+\eta r\cos t}\, r^{p-1}\; \sin^{p-2} t. \quad \ldots \quad (3.3)$$

Expanding $e^{\eta r\cos t}$ in powers of r and integrating out r we get the density function of t as

$$\varphi(t) = \pi^{-\frac{1}{2}}\Big[\Gamma\Big(\frac{p-1}{2}\Big)\Big]^{-1} e^{-\eta^2/2} \sin^{p-2} t \sum_{s=0}^{\infty} \frac{\Gamma[(p+s)/2]}{\Gamma(s+1)}\; (2\eta^2)^{s/2} \cos^s t \quad \ldots \quad (3.4)$$

Substituting in (2.23) we get

$$Pr(uv' < 0) = \pi^{-\frac{1}{2}}\Big[\,\Gamma\Big(\frac{p-1}{2}\Big)\Big]^{-1} e^{-\eta^2/2} \sum_{s=0}^{\infty} \frac{\Gamma[(p+s)/2]}{\Gamma(s+1)}\; (2\eta^2)^{s/2} h_s(\zeta) \quad \ldots \quad (3.5)$$

where
$$h_s(\zeta) = \int_0^{\pi} \sin^{p-2} t \cos^s t \,\Big(\int_{\zeta\cos t}^{\infty} (2\pi)^{-\frac{1}{2}} e^{-x^2/2}\; dx\Big) dt. \qquad \ldots \quad (3.6)$$

We shall obtain a series expansion for $h_s(\zeta)$ valid for all s.

$$h_s(\zeta) = \int_0^\pi \sin^{p-2} t \;\cos^s t\, [(2\pi)^{-\frac{1}{2}} \int_{\zeta \cos t}^{\infty} e^{-x^2/2}\, dx]\, dt$$

$$= \int_0^\pi \sin^{p-2} t \;\cos^s t\, [\tfrac{1}{2} - \int_0^{\zeta \cos t} (2\pi)^{-\frac{1}{2}} e^{-x^2/2}\, dx]\, dt \qquad \dots \quad (3.7)$$

$$= \tfrac{1}{2} \int_0^\pi \sin^{p-2} t \cos^s t\, dt -$$

$$- \frac{1}{(2\pi)^{\frac{1}{2}}} \sum_{k=1} \frac{(-1)^k}{2^k} . \frac{\zeta^{2k+1}}{k\,!} . \frac{1}{2k+1} \int_0^\pi \sin^{p-2} t \cos^{2k+1+s} t\, dt$$

$$= \begin{cases} \frac{1}{2} \quad \dfrac{\Gamma\left(\frac{p-1}{2}\right) \Gamma\left(\frac{s+1}{2}\right)}{\Gamma\left(\frac{p+s}{2}\right)} \;. \quad \text{(if } s \text{ is even)} \\[2ex] -(2\pi)^{-\frac{1}{2}} \displaystyle\sum_{k=0}^{\infty} (-\tfrac{1}{2})^k \frac{\zeta^{2k+1}}{k!(2k+1)} . \frac{\Gamma\left(\frac{p-1}{2}\,\Gamma\right) . \left(\frac{2k+s+2}{2}\right)}{\Gamma\left(\frac{2k+p+s+1}{2}\right)} \end{cases} \qquad \dots \quad (3.8)$$

(if s is odd).

It will not be difficult for the reader to verify that the term by term integration involved is justified.

4. Bounds for the probability

In (3.8) we have a very simple expression for $h_s(\zeta)$ if s is even and an infinite series if s is odd. We shall now determine upper and lower bounds for $h_s(\zeta)$ for s odd. These will in their turn determine upper and lower bounds for $Pr(uv' < 0)$.

If s is odd, we have from (3.7)

$$h_s(\zeta) = -\int_0^\pi \sin^{p-2} t \cos^s t \left(\int_0^{\zeta \cos t} (2\pi)^{-\frac{1}{2}}\; e^{-u^2/2}\; du \right)$$

$$= -2 \int_0^1 (1-x^2)^{\frac{p-3}{2}}\, x^s \left(\int_0^{\zeta x} (2\pi)^{-\frac{1}{2}}\, e^{-u^2/2}\, du \right) dx. \qquad \dots \quad (4.1)$$

Observe that $\int_0^{\zeta x} (2\pi)^{-\frac{1}{2}} e^{-u^2/2}\, du$ is a concave function of x. Hence we have

$$h_s(\zeta) > -\frac{\Gamma\left(\frac{p-1}{2}\right)\Gamma\left(\frac{s+1}{2}\right)}{\Gamma\left(\frac{p+s}{2}\right)} \int_0^{\zeta a(s)} (2\pi)^{-\frac{1}{2}} e^{-u^2/2}\, du \qquad \ldots \quad (4.2)$$

where

$$a(s) = \frac{\Gamma\left(\frac{s+2}{2}\right).\Gamma\left(\frac{p+s}{2}\right)}{\Gamma\left(\frac{s+1}{2}\right).\Gamma\left(\frac{p+s+1}{2}\right)} \qquad \ldots \quad (4.3)$$

Further it is obvious that, for s odd

$$h_s(\zeta) \leqslant 0. \qquad \ldots \quad (4.4)$$

Using (3.8) and (4.2) we obtain from (3.5) the following bounds for $Pr(uv' < 0)$

$$\pi^{-\frac{1}{2}} e^{-\eta^2/2} \left[\frac{1}{2}\sum_{s=0}^{\infty} \frac{\Gamma(s+\frac{1}{2})}{\Gamma(2s+1)} (2\eta^2)^s - \sum_{s=0}^{\infty} \frac{\Gamma(s+1)}{\Gamma(2s+2)} \left(\int_0^{\zeta a(2s+1)} (2\pi)^{-\frac{1}{2}} e^{-u^2/2} du\right) (2\eta^2)^{s+\frac{1}{2}}\right]$$

$$\leqslant Pr(uv' < 0) \leqslant \tfrac{1}{2}\pi^{-\frac{1}{2}} e^{-\eta^2/2} \sum_{s=0}^{\infty} \frac{\Gamma(s+\frac{1}{2})}{\Gamma(2s+1)} (2\eta^2)^s. \qquad \ldots \quad (4.5)$$

These bounds can be written in a simpler form using the duplication formula for the gamma function

$$\pi^{\frac{1}{2}}\Gamma(2x) = 2^{2x-1}\Gamma(x).\ \Gamma(x+\tfrac{1}{2})$$

$$\tfrac{1}{2} - e^{-\eta^2/2} \sum_{s=0}^{\infty} \frac{(\eta^2/2)^{s+\frac{1}{2}}}{\Gamma(s+3/2)} \left(\int_0^{\zeta a(2s+1)} (2\pi)^{-\frac{1}{2}} e^{-u^2/2}\, du\right) \leqslant Pr(uv' < 0) \leqslant \tfrac{1}{2}. \quad \ldots \quad (4.6)$$

These bounds are easier to calculate than the exact expression for $P(uv' < 0)$ itself.

A slightly more simple, but less sharp, lower bound for $Pr(uv' < 0)$ is obtained if we use the fact that, for odd s

$$h_s(\zeta) \leqslant -\frac{\Gamma\left(\frac{p-1}{2}\right).\ \Gamma\left(\frac{s+1}{2}\right)}{\Gamma\left(\frac{p+s}{2}\right)} \int_0^\zeta (2\pi)^{-\frac{1}{2}} e^{-u^2/2}\, du$$

$$\tfrac{1}{2} - e^{-\eta^2/2} \left(\int_0^\zeta (2\pi)^{-\frac{1}{2}} e^{-u^2/2}\, du \right) \left[\sum_{s=0}^{\infty} \frac{(\eta^2/2)^{s+\frac{1}{2}}}{\Gamma(s+3/2)} \right] \leqslant Pr(uv' < 0) \leqslant \tfrac{1}{2} \quad \ldots \text{ (4.7}$$

It is possible to give yet another lower bound. (I am indebted to Dr. C. R. Rao for pointing out this). This is obtained by noting that

$$h_{2s+1}(\zeta) \leqslant |h_{2s+1}(\zeta)| \leqslant h_{2s}(\zeta).$$

Hence we get after some simplification,

$$\tfrac{1}{2} - e^{-\eta^2/2} \sum_{s=0}^{\infty} \frac{(\eta^2/2)^{s+\frac{1}{2}}}{s\,!} \cdot \frac{\Gamma\left(s + \frac{p+1}{2}\right)}{\Gamma(s+p/2)} \leqslant Pr(uv' < 0). \qquad \ldots \quad (4.8)$$

If we separate even and odd terms, the exact formula for $Pr(uv' < 0)$ itself can be expressed more simply as

$$Pr(uv' < 0) = \tfrac{1}{2} + \pi^{-\frac{1}{2}} \left[\Gamma\left(\frac{p-1}{2}\right) \right]^{-1} e^{-\eta^2/2} \sum_{s=0}^{\infty} \frac{\Gamma\left(s + \frac{p+1}{2}\right)}{(2s+1)\,!} (2\eta^2)^{s+\frac{1}{2}} h_{2s+1}(\zeta)$$

$$\ldots \quad (4.9)$$

5. Classification when there are more than two populations

Generalizing the procedure discussed above, we may compute

$$z_i = (N_i/N_i+1)(x^{(i)}-x)\Sigma^{-1}(\bar{x}^{(i)}-x)' \quad (i = 1, 2, \ldots, k) \qquad \ldots \quad (5.1)$$

and assign x to P_j if $z_j < z_i$ $(i = 1, 2, \ldots, k,\ i \neq j)$. Here $\bar{x}^{(i)}$ is the vector of means computed from a sample of size N_i taken from P_i and Σ is the common dispersion matrix assumed to be known. Let us consider the probability of incorrect classification involved in the above procedure. Assume that x really belongs to P_1. We shall give an upper bound to the probability of incorrect classification. Similar upper bounds can be given for the probabilities of misclassification of individuals from $P_2, P_3, \ldots, P_k$.

Let E_i ($i = 2, 3, \ldots, k$) be the event that $z_i < z_1$. Then the chance of misclassifying x is $Pr(\bigcup_{i=2}^{k} E_i)$.

Now, $$Pr\,(\bigcup_{i=2}^{k} E_i) \leq \sum_{i=2}^{k} Pr\,(E_i)\,. \quad \text{(Uspensky, 1937).} \qquad \ldots \quad (5.2)$$

But $Pr(E_i)$ ($i = 2, \ldots, k$) is the probability of misclassification in the situation where there are only two populations P_1 and P_i. We have already explained how this probability can be computed.

References

Fisher, R. A. (1936): The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 7, 179.

Uspensky, J. V. (1937): *Introduction to Mathematical Theory of Probability*, 30, McGraw Hill Book Company Inc., New York and London.

Paper received : November, 1959.