A Note on the Classification of Observation Data

By Hirojiro AOYAMA

(Received April 30, 1950)

1. Introduction. Recently R. v. Mises treated the problem about the classification of data as follows.[1] Let an aggregate be given, upon each element of which a trial is done and consequently one definite real number x is obtained. And let each element of this aggregate belong to one class of n classes which are characterized by n density functions $f_1(x), f_2(x), \cdots, f_n(x)$. Then having a value x as the result of observation, there arises the problem "To which class will this x belong"? As an application of this problem Hayashi recently published an interesting article about the parole prediction.[2] In this note we shall show that we can get similar results as Mises did from some different point of view.

2. Case when there are two classes. We consider now of the certain infinite population where two classes exist, being mixed, which are characterized respectively by the two unimodal density functions $f_1(x)$ and $f_2(x)$ as follows

$$\int_{-\infty}^{\infty} f_1(x)\,dx = 1, \quad \int_{-\infty}^{\infty} f_2(x)\,dx = 1, \tag{1}$$

$$\int_{-\infty}^{\infty} \left\{k_1 f_1(x) + k_2 f_2(x)\right\} dx = 1, \tag{2}$$

k_1 and k_2 being two constants which are proportional rate in the population satisfying $k_1 + k_2 = 1$ and are not necessarily known previously.

When we have an observation data x, we may select the x_0 so as to decide that x should belong to 1st class when $x > x_0$ and to 2nd class when $x < x_0$. Then arises the problem at what value x_0 the reliability will become maximum. Representing the reliability P by the degree of true judging, we may put

$$P = \int_{-\infty}^{x_0} k_2 f_2(x)\,dx + \int_{x_0}^{\infty} k_1 f_1(x)\,dx \tag{3}$$

We may first treat the case when $f_1(x)$ and $f_2(x)$ are both normal density function whose means are m_1 and m_2, and whose standard deviations are σ_1 and σ_2 respectively and when k_1 and k_2 are known previously. For the maximum reliability we get from $\partial \mathrm{P}/\partial x_0 = 0$

$$\frac{k_1}{\sigma_1} e^{-\frac{(x_0-m_1)^2}{2\sigma_1^2}} = \frac{k_2}{\sigma_2} e^{-\frac{(x_0-m_2)^2}{2\sigma_2^2}},$$

that is

$$(\sigma_2^2 - \sigma_1^2)x_1^2 - 2x_1(m_1 \sigma_2^2 - m_2 \sigma_1^2) + \left(m_1^2 \sigma_2^2 - m_2^2 \sigma_1^2 - 2\sigma_1^2 \sigma_2^2 \log \frac{k_1 \sigma_2}{k_2 \sigma_1}\right) = 0. \qquad (4)$$

When $m_1 > m_2$, we may take $x_0 > m_2$ of two roots. When $\sigma_1 = \sigma_2 = \sigma$ we get

$$x_0 = \frac{1}{2}\left\{(m_1 + m_2) - \frac{2\sigma^2}{m_1 - m_2} \log \frac{k_1}{k_2}\right\} \quad . \qquad (5)$$

For every case the reliability is computed from (3). If k_1 and k_2 are not known previously, we decide x_0, k_1 so as to make (3) maximum.

Then putting

$$P= \frac{k_2}{\sqrt{2\pi}\sigma_2}\int_{-\infty}^{x_0} e^{-\frac{(x-m_2)^2}{2\sigma_2^2}} dx + \frac{k_1}{\sqrt{2\pi}\sigma_1}\int_{x_0}^{\infty} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}} dx \qquad (6)$$

where $k_1 + k_2 = 1$, we have from $\partial P/\partial k_1 = 0$

$$\frac{1}{\sigma_1}\int_{x_0}^{\infty} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}} dx = \frac{1}{\sigma_2}\int_{-\infty}^{x_0} e^{-\frac{(x-m_2)^2}{2\sigma_2^2}} dx. \qquad (7)$$

Then from $\partial P/\partial x_0 = 0$ we have also

$$\frac{f_1(x_0)}{f_2(x_0)} = \frac{k_2}{k_1} \quad . \qquad (8)$$

For the general density function we get instead of (7)

$$\int_{x_0}^{\infty} f_1(x)\, dx = \int_{-\infty}^{x} f_2(x)\, dx \quad . \qquad (9)$$

3. Case when there are n classes. Similarly to the previous section we may take n classes which are characterized by the unimodal density functions $f_1(x)$, $f_2(x)$, $\cdots$, $f_n(x)$ respectively and which exist mixed each other in the population. Let the reliability be

$$P = k_1 \int_{-\infty}^{x_1} f_1(x)\, dx + k_2 \int_{x_1}^{x_2} f_2(x)\, dx + \cdots + k_n \int_{x_{n-1}}^{\infty} f_n(x)\, dx \qquad (10)$$

where

$$k_1 + k_2 + \cdots + k_n = 1. \qquad (11)$$

Making P maximum, we have similarly as before

$$\int_{-\infty}^{x_1} f_1(x)\, dx = \int_{x_1}^{x_2} f_2(x)\, dx = \cdots = \int_{x_{n-1}}^{\infty} f_n(x)\, dx \quad . \qquad (12)$$

These $x_1, x_2, \cdots, x_{n-1}$ are the points of division of n classes. And we have at each point of division

$$\frac{f_\nu(x)}{f_{\nu+1}(x)} = \frac{k_{\nu+1}}{k_\nu} \quad (\nu = 1, 2, \cdots, n-1). \tag{13}$$

It we know the value $k_1, k_2, \cdots, k_n$ previously, we can compute directly the reliability from (10). In another case we first decide the x_ν from (12), then decide k_ν from (11) and (13).

4. General case. When n classes are characterized by the m variate unimodal density functions $f_1(x_1, x_2, \cdots, x_m), f_2(x_1, x_2, \cdots, x_m), \cdots, f_n(x_1, x_2, \cdots, x_m)$ resp§tively, we put the reliability P as follows

$$P = k_1 \int_{R_1} f_1\, dR_1 + k_2 \int_{R_2} f_2\, dR_2 + \cdots + k_n \int_{R_n} f_n\, dR_n \tag{14}$$

where $k_1 + k_2 + \cdots + k_n = 1$, and R_ν is a certain region in m-dimensional space, and $R_1 + R_2 + \cdots + R_n$ equal to the whole space. From $\partial P/\partial k_\nu = 0$ we can deduce immediately

$$\int_{R_1} f_1\, dR_1 = \int_{R_2} f_2\, dR_2 = \cdots = \int_{R_n} f_n\, dR_n \tag{15}$$

and if we put the increment ΔP of P equal to zero for the infinitesimal increment ΔR of some region R_μ and the same decrement of adjacent region R_ν, we have

$$\frac{f_\nu(x_1, x_2, \cdots, x_m)}{f_\mu(x_1, x_2, \cdots, x_m)} = \frac{k_\mu}{k_\nu}. \tag{16}$$

So we first decide the regions R_ν from (15) and then compute P from (16) and $k_1 + k_2 + \cdots + k_n = 1$.

References :

1) R. v. Mises: On the classification of observation data into distinct groups, Annals of Mathe. Statis. vol. XVI, No. 1. 1945.

2) Chikio Hayashi: On an application of the statistical method in Parole Prediction. Institute of Case Work.

Institute of Statistical Mathematics
Sangenjaya Laboratory

ERRATA

These Annals, Vol. II, No. 1, 1950. P. 18 insert after last line of section 2

"In these cases we have not the maximum value but only the stationary value just as the minimax solution. If we want to obtain the maximum value, we must estimate the rational rate k_1 and k_2 from experiences in the past time. This fact holds also in the following sections."

Vol. V, No. 1, 1953

Page	line		
27,	9,	read	$M-1-\frac{(M-1)(R-1)}{N}$ instead of the right hand side of (6)
27,	12,	insert	under the assumption after "we have" $N_i=N/R$
27,	14,	read	$2(M-1)+O\left(\frac{1}{N}\right)+O\left(\frac{1}{n}\right)$ instead of the right hand side of (8)
27,	last,	read	(strike off the table)
28,	5-6,	read	(strike off the sentence "under the condition $M=R(R-1)$ and $R \neq 1$"

Vol. VI, No. 1, 1954

Page	line				
13,	12,	read	$\binom{M}{Mp_i}p^{Mp_i}q^{Mq_i}$	instead of	$\binom{M}{Mp_i}p^{Mp_i}p^{Mq_i}$
14,	3,	read	$0.96N$	instead of	$096N$
15,	6,	read	$\ldots k\sqrt{\varepsilon^*D^2(X)}\}\leqq\frac{1}{k^2}$	instead of	$\ldots k\sqrt{\varepsilon^*D^2(X)}\leqq\frac{1}{k^2}$
15,	23,	read	$X_{(i)}$	instead of	$X_{i)}$
24,	7,	read	$-\mu_{11}(2)\mu_{20}(2)\ldots$	instead of	$-\mu_{11}(1)\mu_{20}(2)\ldots$
24,	10,	read	$\frac{N_1^2N_2}{N^3}((\bar{X}_1-\bar{X}_2)\ldots$	instead of	$\frac{N_1^2N_2}{N^3}(\bar{X}_1-\bar{X}_2)\ldots$
25,	9,	read	$\frac{2N_1N_2}{N^3}(\bar{Y}_1-\bar{Y}_2)^2\cdots$	instead of	$\frac{2N_1N_2}{N^2}(\bar{Y}_1-\bar{Y}_2)^2\cdots$
			$+\frac{N_1N_2}{N^5}(N_1^3+N_2^3)\cdots$	instead of	$+\frac{N_1N_2}{N^5}(N_1^2+N_2^3)\cdots$
28,	2 from the bottom,		$+O(n^{-3/2})$	instead of	$+O(^{-3/2})$
30,	11,	read	$-\frac{4\mu_{31}}{\mu_{11}\mu_{20}}-\frac{4\mu_{13}}{\mu_{11}\mu_{02}}+\cdots$	instead of	$-\frac{4\mu_{31}}{\mu_{11}\mu_{21}}-\frac{4\mu_{13}}{\mu_{11}\mu_{12}}+\cdots$
36,	3 from the bottom,		the coming issue	instead of	this issue

Page	line				
54,	6,	**read**	[20], Lemma	**instead of**	[20 , Lemma
68,	28,	**read**	$e^{I\varepsilon(t)}$	**instead of**	$eI^{\varepsilon(t)}$
97,	6,	**read**	(X_j, Y_j) has	**instead of**	$(X_j, Y_j$ has
98,	2,	**read**	$\lim_n \frac{D(S_{nj})}{D_n}$	**instead of**	$\lim_n \sum_l \frac{D(S_{nj})}{D_n}$