RN 002-68

PROBABILISTIC CONCEPT LEARNING

J. de Leeuw

January 1968

PSYCHOLOGICAL INSTITUTE - UNIVERSITY OF LEIDEN - NETHERLANDS

- 1 -

In probabilistic classificatory concept learning tasks (de Klerk & Oppe 1966, Shuford 1964, Lee & Janke 1964, Lee 1963, Lee 1967, de Klerk 1968) concepts may be defined as probability distributions over the stimulus space (de Leeuw 1968b). If the subject's task is to discriminate optimally between the instances of two different concepts, E constructs a set of stimuli by sampling randomly from the two populations associated with these concepts. We shall assume here that the concepts are defined in such a way that:

(1) the two probability distributions are m-dimensional normal distributions.

(2) the dimensions are m independent random variables with equal variances, which implies that the two variance-covariance matrices A_I and A_{II} are identical scalar matrices: $A_I = A_{II} = \text{def}^A$ with every $a_{ii} \neq 0$.

(3) the two vectors of means m_I and m_{II} differ in at least one element: $m_I \neq m_{II}$.

- 2 -

The two multinormal distributions in m dimensions can be written as

$$f_I(y^{(m)}) = (2\pi)^{-\frac{1}{2}m} |A|^{-\frac{1}{2}} \exp. \left\{ -\tfrac{1}{2}(y-m_I)' A^{-1} (y-m_I) \right\} (dy)^m \qquad (1a)$$

$$f_{II}(y^{(m)}) = (2\pi)^{-\frac{1}{2}m} |A|^{-\frac{1}{2}} \exp. \left\{ -\tfrac{1}{2}(y-m_{II})' A^{-1} (y-m_{II}) \right\} (dy)^m \qquad (1b)$$

Now define

$$L(y^{(m)}) = \ln \frac{f_I(y^{(m)})}{f_{II}(y^{(m)})} \qquad (2)$$

Then

$$L(y^{(m)}) = -\tfrac{1}{2} \left\{ (y-m_I)' A^{-1} (y-m_I) - (y-m_{II})' A^{-1} (y-m_{II}) \right\} \qquad (3)$$

This can be simplified in the following way

$$(y-m_I)'A^{-1}(y-m_I) - (y-m_{II})'A^{-1}(y-m_{II}) =$$

$$= (y-m_I)'A^{-1}(y-m_I) - (y-m_{II})'A^{-1}(y-m_I) + (y-m_{II})'A^{-1}(y-m_I) -$$

$$(y-m_{II})'A^{-1}(y-m_{II}) =$$

$$= (m_{II}-m_I)'A^{-1}(y-m_I) + (m_{II}-m_I)'A^{-1}(y-m_{II}) =$$

$$= (m_{II}-m_I)'A^{-1}(2y-(m_I+m_{II})) =$$

$$= 2y'A^{-1}(m_{II}-m_I) - (m_I+m_{II})'A^{-1}(m_{II}-m_I) \qquad (4)$$

From (3) and (4) it follows

$$L(y^{(m)}) = y'A^{-1}(m_I-m_{II}) - \tfrac{1}{2}(m_I+m_{II})'A^{-1}(m_I-m_{II}). \qquad (5)$$

In this way we map the m-dimensional stimuli into the real line. Moreover this loglikelihood axis L is optimal in a discriminant-analytical sense (Van de Geer 1967, Anderson 1958).

- 3 -

In this section we shall prove the following theorem:

The projections of the two multinormal distributions on the loglikelihood axis are distributed as two normal distributions with means $+\frac{1}{2}\alpha$ and $-\frac{1}{2}\alpha$ and with identical standard deviations $\sqrt{\alpha}$, where

$$\alpha = (m_I-m_{II})'A^{-1}(m_I-m_{II}) \qquad (1)$$

Consider the first multinormal distribution: according to assumptions (1) and (2) of section 1, the Y_j $(j=1,\ldots,m)$ are normally distributed random variates with identical standard deviations σ. The characteristic functions are

$$Q_j(t) = \exp(m_j it - \tfrac{1}{2}t^2\sigma^2) \qquad (2)$$

The linear transformation that transforms the vector of coordinates $y_k = (y_{k1},\ldots,y_{km})$ into the corresponding loglikelihood value L_k is given by 2.5:

$$L_k = \frac{m_{I1}-m_{II1}}{\sigma^2} y_{k1} + \cdots + \frac{m_{Im}-m_{IIm}}{\sigma^2} y_{k2}$$
$$-\tfrac{1}{2}(m_I+m_{II})'A^{-1}(m_I-m_{II}) \qquad (3)$$

The first m terms of (3) are still independent normal variates with means $\frac{m_{Ij}-m_{IIj}}{\sigma^2} m_{Ij}$ and variances $\frac{(m_{Ij}-m_{IIj})^2}{\sigma^2}$. This means that their characteristic function can be written as

$$Q_j(t) = \left\{ \exp \frac{m_{Ij}-m_{IIj}}{\sigma^2} m_{Ij} it - \tfrac{1}{2}t^2 \frac{(m_{Ij}-m_{IIj})^2}{\sigma^2} \right\} \qquad (4)$$

The last term in (3) can be interpreted either as a one-point distribution or as a normal distribution with mean $-\frac{1}{2}(m_I+m_{II})'A^{-1}(m_I-m_{II})$ and variance zero. Under both interpretations it is independent of the other variates and its characteristic function is

$$Q_{m+1}(t) = \exp\left\{-\tfrac{1}{2}(m_I+m_{II})'A^{-1}(m_I-m_{II})it\right\} \qquad (5)$$

Because of the independence of the m+1 variates the characteristic function of L can be written as:

$$Q_L(t) = \prod_{j=1}^{m+1} Q_j(t) =$$
$$\exp\left[\left\{\sum_{j=1}^{m}\left(\frac{m_{Ij}-m_{IIj}}{\sigma^2} m_{Ij}\right) - \tfrac{1}{2}(m_I+m_{II})'A^{-1}(m_I-m_{II})\right\} it \right.$$
$$\left. -\tfrac{1}{2}t^2 \sum_{j=1}^{m} \frac{m_{Ij}-m_{IIj})^2}{\sigma^2} \right] \qquad (6)$$

which is again the characteristic function of a normal distribution with mean

$$\mu_1 = m_I'A^{-1}(m_I-m_{II}) - \tfrac{1}{2}(m_I+m_{II})'A^{-1}(m_I-m_{II}) =$$
$$= \tfrac{1}{2}(m_I-m_{II})'A^{-1}(m_I-m_{II}) = \tfrac{1}{2}\alpha \qquad (7)$$

and variance

$$\rho_1^2 = (m_I-m_{II})'A^{-1}(m_I-m_{II}) = \alpha \qquad (8)$$

The same linear transformation can be applied to the second set of Y_j, resulting in a normal distribution with mean $\mu_2=-\frac{1}{2}\alpha$ and variance $\rho_2^2=\rho_1^2=\alpha$. This completes the proof of the theorem.

- 4 -

The α-measure defined in section 3 is identical with Mahanalobis' generalized distance D^2 (Mahanalobis 1936, Anderson 1958, Kendall & Stuart III 1966). From the assumptions in section 1 it follows that A is positive definite and that $m_I-m_{II}\neq 0$. This implies that the quadratic form α is always positive. Moreover, it can be seen that the "standardized" distance between the means of the two groups of projections is equal to

$$\frac{\mu_1-\mu_2}{\rho}=\frac{\frac{1}{2}\alpha+\frac{1}{2}\alpha}{\sqrt{\alpha}}=\sqrt{\alpha}$$

which means that $\sqrt{\alpha}=D$ is identical with the d'-parameter of signal-detection theory (Swets 1964).

- 5 -

We have now characterized all stimuli by a single real number L, the loglikelihood of its belonging to the first of the two populations. We have shown that the projections of the two groups of stimuli on this axis are normally distributed with means $\mu_1=\frac{1}{2}\alpha$ and $\mu_2=-\frac{1}{2}\alpha$ and equal standard deviations $\rho_1=\rho_2=\sqrt{\alpha}$. If we apply another linear transformation to the Y_j and compute the standardized distance between the means of the two resulting groups of projections then $\beta<\sqrt{\alpha}$. This can be seen easily by observing that the axis $L_k=0$ is parallel with Fisher's classical linear discriminant function (Van de Geer 1967).

No proof has been found, but a plausibility argument may be given: For each n=1,2,....

$$\int_{-\infty}^{+\infty} x^n g(x)dx = \int_{-\infty}^{+\infty} x^n f(x)dx \qquad (6)$$

For each n this defines a polynomial function relating the four unknowns η_1, η_2, γ and Q to the parameters of f(x), which are known real numbers. This infinite system of polynomial equations is overdetermined, leaving as the only possible solutions the more or less trivial cases (5a) and (5b).

In words this conjecture amounts to the following:

If the subject's classifications generate two normal distributions with equal standard deviations on the X-axis then either $g_1(x)=f_1(x)$ and $g_2(x)=f_2(x)$ or $g_1(x)=f_2(x)$ and $g_2(x)=f_1(x)$. In both cases the standardized distance between the two distributions is $\sqrt{\alpha}$, so, if the conjecture is true, the only case in which it is formally permitted to compute a d'-parameter is the very special case in which S's classification is exactly identical with (or exactly opposite to) the objective one.

The discussion in section 6 makes it necessary to define a generalisation of d' for the classifications of the subject:

$$\alpha = \frac{\eta_1 - \eta_2}{\sqrt{\frac{1}{2}(\gamma_1^2 + \gamma_2^2)}} \qquad (1)$$

where

$$\eta_i = \int_{-\infty}^{+\infty} x g_i(x)dx \qquad (2)$$

$$\gamma_i^2 = \int_{-\infty}^{+\infty} x^2 g_i(x)dx - \eta_i^2 \qquad (3)$$

- 8 -

A theoretically interesting case is the one prescribed by the optimal strategy in probabilistic concept learning (de Klerk & Oppe, Shuford 1964). This optimal strategy is identical with the classifications rules of statistical decision theory (Anderson 1958, van de Geer 1967). The subject picks out a particular value $\underline{a}$ on the X-continuum and classifies all stimuli with $x > \underline{a}$ as instances of concept I and all others as instances of concept II. The cut-off value $\underline{a}$ for this Ideal Observer (Tanner, Birdsall & Clarke 1960) or Statistical Man (Peterson & Beach 1967) is prescribed by the a priori probabilities and the costs of misclassification. In this paper the general case will be investigated where $\underline{a}$ can have all possible values. To this cut-off strategy with response bias the following applies:

$$g_1(x) = \begin{cases} f(x) & x > \underline{a} \\ 0 & \text{elsewhere} \end{cases} \tag{1}$$

$$g_2(x) = \begin{cases} 0 & x > \underline{a} \\ f(x) & \text{elsewhere} \end{cases} \tag{2}$$

$$P = Q \tag{3}$$

$$\int_{-\infty}^{+\infty} x^n g_1(x)dx = \int_{-\infty}^{+\infty} x^n f(x)dx \tag{4}$$

$$\int_{-\infty}^{+\infty} x^n g_2(x)dx = \int_{-\infty}^{+\infty} x^n f(x)dx \tag{5}$$

The question we want to investigate specifically is:
Given particular values for α, p and $\underline{a}$, what is the α'-value of 7.1? The special case with $\underline{a}=0$ has been studied by Van de Geer (personal communication). This special case corresponds to the Ideal Observer strategy with equal a priori probabilities and equal costs of misclassification. It is easy to show that this $\underline{a}=0$ case is the only case where $\gamma_1=\gamma_2$. The actual integrations

and computations are presented in a companion paper (de Leeuw, 1968c).

- 9 -

In probabilistic scalar concept learning the subject's task is to give numerical responses, his subjective probabilities $\psi(H_1/D)$ that the particular stimulus is a positive instance of the concept. The model for this kind of experiments postulates that these subjective probabilities result from the applications of Bayes' Rule to two subjective $\psi(D/H_i)$ densities. If we postulate in addition that these subjective distributions are normal with equal variances, it follows easily that the subjective loglikelihood ratio

$$SLLR = \ln \frac{\psi(H_1/D)}{1-\psi(H_1/D)} = \lambda \; BLLR + \mu$$

is a linear function of the objective loglikelihood ratio, defined in section 2 of this paper (also cf. de Leeuw, 1968a.)

Theorem: If the above mentioned assumptions are met, the standardized distance between the two subjective distributions is equal to the standardized distance between the objective distributions.

Proof: For the stimuli of the first set

$$\mathcal{E}(SLLR_I) = \lambda \mathcal{E}(BLLR_I) + \mu = -\tfrac{1}{2}\alpha\lambda + \mu$$

In the same way for the second set

$$\mathcal{E}(SLLR_{II}) = \lambda \mathcal{E}(BLLR_{II}) + \mu = \tfrac{1}{2}\alpha\lambda + \mu$$

For the variances we obtain

$$VAR(SLLR_I) = \mathcal{E}(SLLR_I^2) - (\mathcal{E}(SLLR_I))^2 = \lambda^2\alpha$$

which implies, of course,

$$VAR(SLLR_{II}) = \lambda^2\alpha$$

This means that the standardized distance is equal to $\sqrt{\alpha}$. Q.E.D.

Assuming equal a priori probabilities ($P=\frac{1}{2}$) we obtain for the total set of stimuli

$$\mathcal{E}(SL^{L}R) = \mu$$

while

$$\mathcal{E}(BLLR) = 0$$

and

$$VAR(SLLR) = \lambda^{2}\alpha$$

while

$$VAR(BLLR) = \alpha$$

The interpretation of these parameters is clear: μ is a bias-parameter and λ^{2} reflects the degree of conservatism in the subjective probabilities.

LITERATURE

Anderson, T.W. : An introduction to Multivariate Statistical Analysis, New York Wiley, 1958.

De Klerk, L.F.W. & Oppe, S. : Probabilistic Concept Learning, Report E 013-66, Psychological Institute, University of Leiden, 1966.

De Klerk, L.F.W. : Probabilistic Concept Learning, Doctoral Dissertation (in press), University of Leiden, 1968.

De Leeuw, J. : De relatie tussen subjektieve en Bayesiaanse a posteriori waarschijnlijkheden, Unpublished paper, Psychological Institute, University of Leiden, 1968a.

De Leeuw, J. : Notes on the definition of a concept, Research Note RN 001-68, Psychological Institute, University of Leiden, 1968b.

De Leeuw, J. : The effect of a cut-off strategy on the alpha prime measure, Research Note RN 003-68, Psychological Institute, University of Leiden, 1968c.

Kendall, M.G. & Stuart, A. : The advanced theory of statistics, vol. III, London, Griffin, 1966.

Lee, W. : Choosing among confusably distributed stimuli with specified likelihood ratio's, Perceptual, Motor and Skills, 1963, 6, 445-467.

Lee, W. : Conditioning parameter model for reinforcement generalisation in probabilistic discrimination learning; J. Math. Psychol., 1966, 3, 184-196.

Lee, W & Janke, M. : Categorizing externally distributed stimulus samples for three continua; J. Exp. Psychol., 1964, 68, 376-382.

Peterson, C.R. & Beach, L.R. : Man as a intuitive statistician Psychol. Bull., 1967, 68, 29-46.

Mahalanobis, P.C. : On the generalized distance in statistics, Proc. Nat. Inst. Sci. (India), 1936, 12, 49-55.

Shuford, E.H. : Some Bayesian learning processes, in M.W. Shelley III & G.L. Bryan (eds) Human judgements and optimality, New York Wiley, 1964.

Swets, J.A. (ed) : Signal detection and recognition by human observers, New York Wiley, 1964.

Tanner, W.P. Jr., Birdsall, T.G. & Clarke, F.R. : The concept of the ideal observer in psychophysics, T.R. 98, Electronic Defense Group, University of Michigan, 1960.

Van de Geer, J.P. : Inleiding in de multivariate Analyse, Arnhem Van Loghum Slaterus, 1967.