Ελληνικό Στατιστικό Ινστιτούτο
Πρακτικά 18ου Πανελληνίου Συνεδρίου Στατιστικής (2005) σελ.419-426

CLASSIFICATION UNDER A MULTIVARIATE BERNOULLI: AN APPLICATION TO PNEUMOCONIOSIS

T. Cacoullos and M. Pattichis
University of Athens and University of New Mexico

SUMMARY

Two training samples radiographs of 158 pneumoconiosis subjects, allocated in one of two categories, led to the classification problem for two populations under a six-dimensional Bernoulli distribution. Five classifiers: logistic regression, Bayes, *k*-means, simple and weighted sums, are considered; their apparent misclassification errors are evaluated; they range from 9% to 12%; the logistic has the smallest error, 9%.

1. INTRODUCTION

The discrimination problem below for a multivariate Bernoulli distribution arose in relation to the classificatin of chest radio-graphs of a number of subjects (pneumoconiosis patients – miners in New Mexico) into one of three *q*-categories q_0, q_1, q_2 (according to the ILO, International Labour Office, categorization into main categories p, q, r, s, t, u)[*]. More specifically, out of $n = 158$ subjects, 126 were classified as q_0, 26 as q_1 and only 6 as q_2; the rater (radiologist) based his *q*-classification on his observing the presence (1) or absence (0) of a mark (e.g., opacity) in each of the six regions $R_1, ..., R_6$, into which the two lungs were divided for observational purposes; R_1, R_2, R_3 for the right lung and R_4, R_5, R_6 for the left lung (see Figure 1); R_1 and R_4 are the two upper regions, R_2 and R_5 are the middle ones, and R_3 and R_6 the lower ones.

[*] p (in particular p_0) denotes normals, q the next to normal, etc (in increasing order of gravity)

Thus for describing the regional (spatial) lung variation, we define Bernoulli random variables (rv) $X_i = 1$ or 0 according to the appearance (1) or not (0) of a mark in each region R_i; hence each radiograph gives rise to a six-dimensional Bernoulli rv

	Right	Left
Upper	$R_1(UR)$	$R_4(UL)$
Middle	$R_2(MR)$	$R_5(ML)$
Lower	$R_3(LR)$	$R_6(LL)$

$\mathbf{X} = (X_1, ..., X_6)$. For example, $\mathbf{x} = (1, 0, 0, 1, 0, 0)$ means marks only in the upper parts of the lungs (the middle and lower parts are "clear").

2. SOLVING THE DISCRIMINATION PROBLEM

The preceding introduction leads to the following statistical discrimination problem: Given a training sample of n subjects classifed (by a doctor) into one of the q-subcategories q_0, q_1, q_2, what is a reasonable (statistical) rule for allocating (classifying) a new subject (radiograph) $\mathbf{x} = (x_1, ..., x_6)$ to one of the categories (populations) q_0, q_1, q_2? In the present pneumoconiosis example, the sample sizes $n_0 = 126$, $n_1 = 26$ and $n_2 = 6$, from q_0, q_1 and q_2, are actually random variables obtained from the mixutre of $n = n_0 + n_1 + n_2 = 126 + 26 + 6 = 158$ chest radiographs, a sample from a large population of q-type patterms of pneumoconiosis. However, here we will apply the logistic and other discrimination procedures as if separate training samples were given from the q-categories, in view of the smalleness of the sample sizes and the estimation problems for mixtures (for a discussion of these matters see Anderson, 1972, 1973). Furthermore, only $n = 6$ patterns (in fact very little differing from q_1-patterns) were allocated to q_2, making it impossible to estimate the parameters involved in the discrimination problem; hence, we pooled the n_2 q_2-patterns with the $n_1 = 26$ q_1 patterns, thus arriving at a 2-category discrimination problem with $n_0 = 126$ q_0-patterns and $n_1' = n_1 + n_2 = 26 + 6 = 32$ q_1-patterns.

The preceding reduction to a two-category discrimination problem, in addition to the smallness of $n_2 = 6$, takes also into account the fact that the general optimum Bayes discriminatory rule, with respect to prior probabilities P_0 for q_0 and P_1 for q_1 $(P_0 + P_1 = 1)$, allocates a sample point $\mathbf{x}$ to q_0 if

$$P_0 p_0(\mathbf{x}) \geq P_1 p_1(\mathbf{x}) \quad \text{or} \quad P[q_0 \mid \mathbf{x}] \geq P[q_1 \mid \mathbf{x}] \tag{1}$$

(and to q_1, otherwise), where $p_j(\mathbf{x})$ is the probability function of $\mathbf{x}$ under $q_j(\mathbf{x})$ $(j = 0,1)$ which in the present 6-dimensional Bernoulli case involves $2^6 = 64$ unknown probabilities, the probabilities

$$p(\mathbf{x}) = P[X_i = x_i, i = 1,...,6], \text{ with } x_i = 0 \quad \text{or} \quad 1, \quad i = 1,...,6;$$

$P(q_j \mid x)$ denotes the posterior probability of q_j given $\mathbf{x}$. Obviously much larger (training) samples n_0 and n_1' are necessary for estimating 63 parameters (since $\sum_{\mathbf{x}} p(\mathbf{x}) = 1$), though in the present case one q_0 pattern, the $(0,0,0,0,0,0)$ has frequency 76% (should perhaps be classified into the *p*-category, i.e., normals) under q_0 and another 10 or so with a frequency of about 1% (see Table 3.1 below). In fact, of the 64 possible patterns $\mathbf{x} = (x_1,...,x_6)$, only 11 different ones appeared. Nonetheless, the estimation problem of the $p_j(\mathbf{x})$ does not become any easier, since the (maximum likelihood) estimates of $p(\mathbf{x})$ for the missing patterns (with zero frequencies) are zero (of course, not true). Another consequence of this is that it is not possible to classify a future $\mathbf{x}$ which has not appeared in the training sample, at least by Bayes rules, which involve the (unknown) $p_j(\mathbf{x})$.

The preceding considerations point out the difficulties in the treatment of the present discrete multivariate discrimination problem, especially for Bayes discrimination rules in view of (1), where the $p_j(\mathbf{x})$ have to be estimated from the available small training samples.

In the sequel, in addition to the logistic regression (discrimination) approach and the (plucg-in) Bayes and *k*-means rules, we use simple ad hoc classifiers, such as the simple sum $s = x_1 + \cdots + x_6$ or a weighted sum s^* (see (11) below) of the x_i, in view of the fact that pneumoconiosis was shown to start in the upper lungs ($x_1 = 1$, $x_4 = 1$, say) and progresses (more 1's) downwards (q_2-patterns have, on average, more $x_i = 1$; see Table 1 and 5 below).

3. SEVERAL CLASSIFIERS AND THEIR PERFORMANCE

In this section, we give the classifiers, as motivated in the preceding section, along with their coresponding Apparent (as estimated from the training samples) Total Misclassification (probabilities) Error (ATME); this is simply the percentage of wrongly classified patterns of the training samples.
The training samples are summarized in the following

Table: *The* $\mathbf{q_0}$ *and* $\mathbf{q_1}$ *patterns*

q_0 patterns (126)			q_1 patterns (32)		
Pattern	Frequency	Percent	Pattern	Frequency	Percent
000-000	96/126	76%	000-100	1/32	3%
000-100	1/126	1%	100-100	13/32	41%
001-101	1/126	1%	100-110	1/32	3%
100-000	6/126	5%	110-100	1/32	3%
100-100	18/126	14%	110-110	10/32	31%
101-001	1/126	1%			
110-100	1/126	1%	111-110	1/32	3%
110-110	1/126	1%	111-111	5/32	16%
111-111	1/126	1%			

Note the big overlap of q_0 and q_1 at the pattern 100-100, which tends to increase the misclassification errors.

A. Logistic Discrimination. In the logistic form for the posterior probabilities for $s = 2$ categories in the well-known Cox-Day-Kerridge approach the fitted $y = 0$ or 1, in the form

$$E(y) = \frac{1}{1+e^{\boldsymbol{\alpha}'\mathbf{x}}} \equiv \pi(\mathbf{x}) = \pi \quad \text{logit}\, y = \log\frac{\pi}{1-\pi} = \boldsymbol{\alpha}'\mathbf{x} = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_s x_s \qquad (2)$$

In the present application, with $\mathbf{x} = (x_1,...,x_6)$, $x_i = 0$ or 1, $s = 6$.

Using the q_0 and q_1 training samples from q_0 and q_1, a logistic regression program (see, e.g., Hosmer and Lemershow, 2000) gave us the logit in (2) with $s = 6$ equal to

$$5.547 + 0.187x_1 - 1.925x_2 - 4.593x_3 + 0.187x_4 - 0{,}966x_5 - 1.079x_6.$$

If d_0 denotes allocation of an **x** to q_0 and d_1 to q_1, then we get the logistic regression (logit) classifier of Table A. That is, out of the 126 q_0 patterns the classifier allocated 114 to q_0 (correctly) and 12 to q_1 (incorrectly); whereas out of the 32 q_1 patterns 2 were (incorrectly) allocated to q_0 and 30 (correctly) to q_1. Hence the corresponding apparent totat error

***Table A** Logit Classifier*

	d_0	d_1	total
q_0	114	12	126
q_1	2	30	32
total	116	32	158

$$\text{ATME} = \frac{12+2}{158} = 0.09. \qquad \textbf{(3)}$$

B. The Bayes Classifier. The Bayes classifier (cf. (1)) gave the following Table B.

***Table B.** Bayes classifier*

	d_0	d_1	total
q_0	125	1	126
q_1	15	17	32
total	140	18	158

The Bayes classifier does better than the logit in classifying the q_0 patterns (125 of the 126) whereas the logit is better in classifying the q_1 patterns (30 of the 32). The

$$\text{Bayes classifier ATME} = \frac{1+15}{158} = 0{,}10, \qquad \textbf{(4)}$$

slightly higher than the logit, with $\text{ATME} = 0.09$ of (3).

C. The k-means classification ($k = 2$). This classifier allocated an $\mathbf{x} = (x_1, ..., x_6)$ to the q category whose mean is closer (in Euclidean distance) to $\mathbf{x}$. The q-mean vectors $\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1$ are:

For q_0, $\bar{\mathbf{x}}_0 = (0.643, 0.286, 0.214, 0.571, 0.143, 0.214)$ and **(5)**

for q_1, $\bar{\mathbf{x}}_1 = (0.824, 0.529, 0.235, 0.941, 0.529, 0.235)$.

Table C. *Two means* **classifier**

	q_0	q_1	total
d_0	124	14	12
d_1	2	18	20
total	126	32	158

The 2-means allocation is shown in Table C. Its

$$\text{ATME} = \frac{14+2}{158} = 0.10, \qquad \textbf{(6)}$$

equal to the Bayes ATME in (4).

D. A simple sum classifier. It is easily observed (see Table 1) that marks $(x_i\text{'}s\ 1)$ start appearing at the upper lung regions so that pneumoconiosis starts at the upper lungs and progresses downwards. Thus a reasonable and very simple, ad hoc, classifier of an $\mathbf{x} = (x_1, ..., x_6)$ can be based on the number of marks

$$s = s(\mathbf{x}) = x_1 + x_2 + \cdots + x_6, \quad s = 0, 1, ..., 6. \qquad \textbf{(7)}$$

We evaluated the classifier:

Allocate $\mathbf{x}$ to q_0 if $0 \le s \le 1$; to q_1 if $2 \le s \le 6$. The corresponding allocation, Table D, gave a total error

$$\text{ATME} = \frac{16+3}{158} = 0.12$$

Table D. *Sum classifier*

	q_0	q_1	total
d_0	110	3	113
d_1	16	29	45
total	126	32	158

E. A weighted-Sum classifier. A more sophisticated classifier can be based on assigning bigger weights to x_i's corresponding to the middle and lower lung regions. This is accomplished by introducing some new random variables based on the 6 x_i's. This was motivated by the representation, Teugels (1990), of an *s*-variate Bernoulli rv $\mathbf{x} = (x_1,...,x_s)$ with probabilities

$$p(\mathbf{x}) = P[X_1 = x_1,...,X_s = x_s], \quad x_i = 0,1 \quad (i = 1,...,s)$$

in terms of a one-dimensional rv $\xi = \xi(\mathbf{x})$, taking 2^s values and defined by

$$\xi = \xi(x) = 1 + \sum_{i=1}^{s} 2^{i-1} x_i \,, \quad \xi = 1,2,...,2^s . \tag{8}$$

Indeed, there is a one-to-one correspondence between the $p(\mathbf{x})$ probabilities and the 2^s probabilities

$$P[\xi(\mathbf{X}) = \xi(\mathbf{x})].$$

Here $s = 6$ and we define a ξ transform for the right-lung triplet (x_1, x_2, x_3) $(s = 3)$ namely,

$$\xi_1 \equiv \xi_1(x_1, x_2, x_3) = 1 + \sum_{i=1}^{3} 2^{i-1} x_i = 1 + x_1 + 2x_2 + 4x_3 \,, \quad \xi_1 = 1,....,8 \,, \tag{9}$$

and similarly for the left-lung triplet (x_4, x_5, x_6)

$$\xi_2 \equiv \xi_2(x_4, x_5, x_6) = 1 + x_4 + 2x_5 + 4x_6 \,, \quad \xi_2 = 1,....,8 \,, \tag{10}$$

so that the resulting sum $\xi_1 + \xi_2$ takes the 15 values $2,...,16$. Moreover, the sum, say s^*, of ξ_1 and ξ_2,

$$s^* = \xi_1 + \xi_2 \,, \tag{11}$$

is expected to give a highter discriminatory power than the simple sum s of 6, providing more choices for the separating value, s_0^*, say, of s^* for the two categories, that is, allocate to q_0 if $s^* \le s_0^*$, otherwise allocate to q_1. Moreover, s^*

is a weighted sum of the 6 x_i's, giving more weight as we move from the upper to the middle and the lower parts of the lungs, as shown by (9) and (10).
In view of the preceding remarks, we considered and evaluated the following classifier, based either on the sum $S^* = \xi_1 + \xi_2$ or one of the ξ_1, ξ_2, namely, allocate to q_1 if either $\xi_1 + \xi_2 > 4$ or $\xi_1 > 2$ or $\xi_2 > 2$; otherwise, to q_0.

This weighted-sum classifier gave the classification of Table E.

Table E. *Weighted sum*

	q_0	q_1	total
d_0	122	14	136
d_1	4	18	22
total	126	32	158

The misclassification error shows a slight improvement compared to the simple-sum *s* error.

Its $\text{ATME} = \frac{4+14}{158} = 0{,}11$

Remark.

It is observed that all 5 classifiers considered above have comparable discriminatory powers, as measured by the apparent total misclassification error ATME, ranging from 9% for the logistic to 12% for the simple ad hoc sum classifier *s* of (7).
In conclusion, the present application of the logistic model to a multivariate Bernoulli situation provides another example of its usefulness and good performance, especially in treating discrete data. It also works for case of more than two categories as well as for multinomial situations (see McCullagh and Nelder, 1989).

ΠΕΡΙΛΗΨΗ

Μία ακτινογραφία των πνευμόνων πάσχοντος από πνευμονοκονίαση δίνει μίαν εξαδιάστατη μεταβλητή Bernoulli. Με βάση τα διδακτικά (training) δείγματα 126 ατόμων της κατηγορίας (βαθμού πνευμονοκονίασης) q_0 και 32 της q_1, προτείνονται πέντε ταξινομικοί κανόνες: ο «λογιστικός» (logistic), Bayes, *k*-means, απλού αθροίσματος και σταθμισμένου αθροίσματος. Το εμπειρικό σφάλμα ταξινόμησης κυμαίνεται από 9% έως 12%, με ελάχιστο της logistic.

REFERENCES

Andesron, J.A. (1972) Separate sample logistic discrimination, *Biometrika*, **59**, 19-35

Anderson, J.A. (1973) Logistic discrimination with medical applications, in *Discriminant Analysis and Applications* (T. Cacoullos, ed.), Academic Press, New York, 1973.

Hosmer, D. and Lemershow, S. *Applications of Logistic Regression*, Wiley, New York, 2000.

McCullagh, P. and Nelder, J.A. *Generalized Linear Models*, Chapman & Hall, London, 1989.

Teugels, J.L. (1990). Some representations of the multivariate Bernoulli and binomial distributions. *Journal of Multivariate Analysis*, **32**, 256-268.