The Statistician (1987) 36, pp. 107–113

Classification problems in education

J. D. BERMUDEZ, J. M. BERNARDO & M. SENDRA

Departamento de Estadistica, Facultad de Matimaticas, Universidad de Valencia, 46071 Valencia, Spain

Abstract. The paper suggests that Bayesian probabilistic classification provides an interesting framework to analyse the data banks typically encountered in Education research. Particular procedures are suggested to analyse the relationship between interesting partitions of a student population and the student profiles. A case study is described as an example.

1 Introduction

Education research is often based in the analysis of data banks which consist of the measurements of a miscellaneous collection of attributes on certain class of students. The attributes, or *items*, which constitute the *profile* of each student may provide information about his or her physical characteristics (age, sex,...) sociological data (parents' education, number of brothers and sisters, family habitat,..) social attitudes towards key issues (sex or race discrimination, politics,..) test results (abstract reasoning, IQ tests, reading comprehension, spelling,..) and academic grades in different subjects (mathematics, english, history,..). With this type of information the researcher tries to answer different types of questions such as (i) may a certain sample of the students in the data bank (say those taken as controls in a given study) be considered exchangeable with the rest? (ii) is there any (statistical) relationship between a group of items (say sociological profile) and the result of a test (say on mathematical ability)?.

We argue that a useful solution to these and many other interesting questions may be provided within the framework of (Bayesian) probabilistic classification, where the data bank is categorised into a finite number of interesting classes $\{\delta_1, \delta_2,\ldots, \delta_k\}$ using one or more of the items measured and then the predictive probability distribution $\{p(\delta_1|\mathbf{x},D),\ldots, p(\delta_k|\mathbf{x},D)\}$ of a student with profile $\mathbf{x}$ belonging to each of the possible classes given the data bank D is computed in order to assess the dependence of the classification considered on the students' profiles.

Thus, in an experimental setting students may be categorised into either *controls* (δ_1) or *experimentals* (δ_2); derivation of the predictive probability distribution $\{p(\delta_1|\mathbf{x},D), p(\delta_2|\mathbf{x},D)\}$ then provides the relevant information on whether or not controls and experimentals students are similar. Indeed, if for some $\mathbf{x}$ this distribution is far from (1/2, 1/2), students of the type described by that $\mathbf{x}$ are not evenly distributed among the two categories. Similarly, if the data bank is categorised into, say, those who excel (δ_1) pass (δ_2) or fail (δ_3) a mathematics test, the predictive distribution $\{p(\delta_1|\mathbf{x},D), p(\delta_2|\mathbf{x},D), p(\delta_3|\mathbf{x},D)\}$ may be used to identify, say, those students who may be expected to do well in mathematics: those whose profile $\mathbf{x}$ makes $p(\delta_1|\mathbf{x},D)$ large.

2 The classification model

An important advantage of a classification approach over more standard statistical analyses of these problems lies in the fact that it may be carried out with the large

number of highly interdependent attributes of any type (continuous, binary, ordinal,..) which typically form the student profile; the model which follows was proposed in Bernardo (1987).

Consider the linear discriminant function

$$\lambda_i^t \mathbf{x} = (x_{1.},\ldots,x_{k.})^t \, S^{-1} \, \mathbf{x}, \qquad i=1,\ldots,k-1$$

where $x_{i.}$ is the mean of the profiles in class δ_i and S their pooled covariance matrix. It is well known that $\lambda_i^t x$ is the linear function which best separates δ_i from δ_k (Fisher, 1936; Goel, 1983). Suppose, without loss of generality, that the x's represent *standardised* quantities. We shall assume:

A1 (Approximate sufficiency)

$$p(\delta_k|\mathbf{x},D) \approx p(\delta_k|t,D) \qquad \boldsymbol{t}=\boldsymbol{t}(\mathbf{x})=\{\lambda'_1\mathbf{x},\ldots,\lambda'_{k-1}\mathbf{x}\}$$

The assumption of the approximate sufficiency for classification purposes of the linear discriminant functions is empirically founded in large field studies (see, e.g. Titterington *et al.*, 1981) and may often be regarded as a good first order approximation. This assumption dramatically reduces the dimensionality of the problem.

A2 (Approximate normality)

$$p(\boldsymbol{t}|\delta_i) \cong N_{k-1}(\boldsymbol{t}|\mu_i,\Sigma_i), \; i=1,2,\ldots,k$$

The assumption of approximate joint normality of the sampling distribution of the linear functions $\mathbf{t}$ is founded on central limit type arguments (see, e.g. Diaconis & Freedman, 1984) for, each $\boldsymbol{t}$ is the sum of a large number of standardised random quantities neither of which is expected to dominate the others. Note that this assumption will *not* be sensible on the sampling distribution of the profile $\mathbf{x}$.

A3 (Reference prior)

$$p(\mu_i,\Sigma_i|\delta_i) \propto |\Sigma_i|^{-k/2}$$

The prior information about the values of the discriminant function is typically very vague: thus, it may well be approximated by the appropriate reference *non-informative* prior (Bernardo, 1979b). It follows from A1 to A3 (see, e.g. Geisser, 1964) that the predictive distribution is of the form

$$p(\delta_i|\mathbf{x},D) \propto St(\boldsymbol{t}|\boldsymbol{m}_i,(n_i+1)/(n_i-k+1)\boldsymbol{V}_i, \; n_i-k+1)p(\delta_i|D)$$

where $\boldsymbol{m}_i$ and $\boldsymbol{V}_i$ are respectively the mean vector and covariance matrix of the $\boldsymbol{t}$ vector *with* class δ_i, n_i is the number of students in class δ_i, $St(\boldsymbol{t}|\mathbf{m},\boldsymbol{V},\alpha)$ is a $(k-1)$ variate Student density with mean $\boldsymbol{m}$, dispersion matrix $\boldsymbol{V}$ and α degrees of freedom and $p(\delta_i|D)$ is either the prior probability $p(\delta_i)$ of category δ_i (with retrospective sampling) or its (unconditional) posterior probability $(n_i+1/2)/(n+k/2)$, $n=\Sigma\, n_i$, (with prospective sampling).

3 The choice of variables

The selection of an appropriate subset of attributes $\boldsymbol{x}$ is important not just because it reduces computing but mainly because it serves to identify those attributes within the student profiles which are essential in each classification problem; the procedure which follows was proposed in Bernardo and Bermudez (1985).

Consider the selection of attributes as the decision problem described in Fig. 1, so that the statistician selects a subset of attributes, observes its values on a randomly

chosen student and gets a reward $u\{p(\delta|\mathbf{x}_f,D)\}$ which depends on the (predictive) probability produced for the category δ which actually obtains.

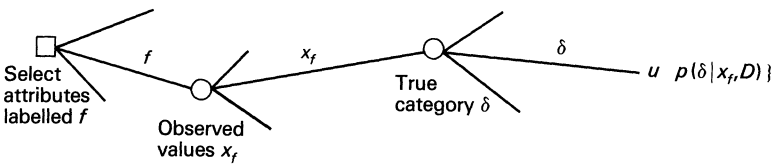


Fig. 1. Diagram showing the decision tree.

It is well known (Savage, 1971; Bernardo, 1979a; Lindley, 1982) that u should be required to be a *proper* scoring rule and that the only proper scoring rules which only depends on the predictive distribution through the probability associated to the true category is a *logarithmic* scoring rule of the form

$$u\{p(\delta|\mathbf{x},D)\}=A\log p(\delta|\mathbf{x},D)+B(\delta),\quad A>0$$

It follows that the best subset of attributes is that which maximises

$$u^*\{f\}=\int p(\mathbf{x}_f|D)\sum_{i=1}^{k}p(\delta_i|\mathbf{x}_f,D)\log p(\delta_i|\mathbf{x}_f,D)\,d\mathbf{x}_f$$

i.e. that which minimises the expected value of the entropy of the resulting predictive distribution. Using Monte Carlo methods, this integral may be approximated by

$$\frac{1}{n}\sum_{j=1}^{n}p(\delta_i|\mathbf{x}_{fj},D)\log p(\delta_i|\mathbf{x}_{fj},D)$$

with prospective sampling or

$$\sum_{l=1}^{k}p(\delta_l)\frac{1}{n_l}\sum_{j=1}^{nl}\sum_{i=1}^{k}p(\delta_i|\mathbf{x}_{fj},D)\log p(\delta_i|\mathbf{x}_{fj},D)$$

with retrospective sampling. Thus the optimal solution approximately consists in choosing that subset of attributes which minimises the average entropy of the predictive distribution.

A complete search among all possible subsets of attributes is computationally unfeasible. Thus, we propose a sequential algorithm which, given an initial subset (possibly empty) (i) sequentially eliminates attributes while the associated expected utility does not decrease more than some constant ε in each step, thus eliminating possible redundancies within the initial subset and (ii) sequentially incorporates new attributes while the associated expected utility increases at least ε in each step.

4 The predictive distribution of the classification probabilities

The use of a classification approach to the problems discussed provides a very rich answer in that it gives a solution $\{p(\delta_i|\mathbf{x},D)\ i=1,2,\ldots,k\}$ as a function of the student profile $\mathbf{x}$. It is however important to be able to *summarise* the results obtained with a description of the classification probabilities $\{p(\delta_i|\mathbf{x},D)\ i=1,2,\ldots,k\}$ which obtain as $\mathbf{x}$ ranges over the class of possible profiles.

The natural answer is to produce the (predictive) probability distribution of the classification probabilities $p(\delta_i|\mathbf{x},D)=p(\delta_i|\mathbf{t},D)$, $i=1,2,\ldots,k$. Since those are well-defined

mathematical functions of the random vector $\boldsymbol{t}$, the problem reduces to compute the probability distribution of the function $y_i(\boldsymbol{t})=p(\delta_i|\boldsymbol{t},D)$ where, as described in Section 2,

$$p(\boldsymbol{t}|\delta_i)=St(\boldsymbol{t}|\boldsymbol{m}_i,(n_i+1)/(n_i-k+1)\boldsymbol{V}_i,\ n_i-k+1)\cong N_{k-1}(\boldsymbol{t}|\boldsymbol{m}_i,\boldsymbol{V}_i)$$

We do not have a general expression for the probability function of $\boldsymbol{y}(\boldsymbol{t})$, although it is always possible to obtain particular solutions. As an example, we present here the solution for the very important case $k=2$, $V_1\neq V_2$ (without loss of generality it is always possible to assume $V_1<V_2$) and $p(\delta_1)=p(\delta_2)=1/2$.

For that particular case, the function $y_1(t)=p(\delta_1|t,D)$ reduces to:

$$y=y_1(t)=\frac{1}{1+\exp\{A+B(t-M)^2\}}$$

where

$$A=-\tfrac{1}{2}\text{Log}\,\frac{V_2}{V_1}-\frac{(m_1-m_2)^2}{2(V_2-V_1)},\ B=\frac{V_2-V_1}{2V_1V_2},\ M=\frac{V_2\,m_1-V_1\,m_2}{V_2-V_1},$$

and the predictive distribution of t is given by:

$$p(t)=1/2\{N(t|m_1,V_1)+N(t|m_2,V_2)\}.$$

Then, some tedious, but straightforward, algebra permit to prove the following

Proposition 1

With the above notation, the (predictive) density function of y is given by:

$$p(y)=\frac{N(M+R(y)|m_1,V_1)+N(M-R(y)|m_1,V_1)+N(M+R(y)|m_2,V_2)+N(M-R(y)|m_2,V_2)}{4\,R(t)\,B\,y\,(1-y)}$$

if y belongs to the interval $(0,\{1+\exp(A)\}^{-1})$, $p(y)=0$ otherwise. Where

$$R(y)=\sqrt{\frac{\log((1-y)/y)-A}{B}}$$

5 A case study

The Spanish socialist government is currently preparing a large reform of the structure of secondary education. To gather information which could be used to improve the education system and to test the consequences of some proposed reforms before being generally implemented, 23 196 students were selected in 262 schools and subjected to a large number of tests. The profile of each of those 23 196 students finally contained over 150 items, including their sociological description, social attitudes, results to different types of tests and academic grades in former education. We were requested to analyse those data.

Example 1

The education authorities showed interest in analysing the factors which motivated in Spain the choice of a private rather than a public (state-owned) school. Thus, we categorised the data bank into students attending a private (d_1) or a public (d_2) school, and used the methodology just described (i) to determine the relevant classificatory

variables and (ii) to obtain the predictive distribution of the classification probabilities.

The variables sequentially selected and their corresponding λ-coefficients were:

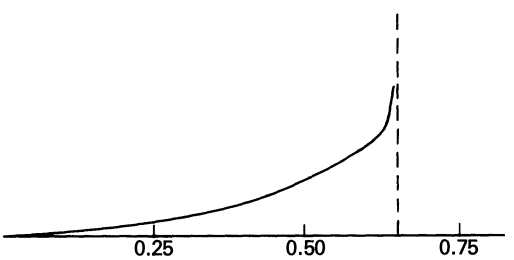
Variables	λ_i
Education level of father	0·594
Nursery school attended	0·313
Proximity house-school	0·301
Positive opinion on politics	−0·267
Number of brothers/sisters	−0·241
Abstract reasoning	0·229
Math. computation ability	0·132
Reading comprehension	−0·127
Mother works fulltime	−0·121
Critical power	0·095
Lost academic years	−0·046

For each individual we computed the values of the linear discriminant function $t=\lambda'\mathbf{x}$. The Table below summarises the corresponding results within each category group:

	Number Obs.	*Mean*	*Std. Dev.*	*Min.*	*Max.*
Public School	9587	−0·238	0·932	−3·523	3·748
Private School	7725	0·292	1·078	−3·188	4·077

The predictive distribution of the probability $p(\delta_1|\mathbf{x},D)$ that a random student belongs to a public school was found to be

Total Number Obs.	17312
Mean	0·500
Std. Dev.	0·134
Minimum	0·000
Maximum	0·651
50% H.P.D.	(0·422 , 0·605)
90% H.P.D.	(0·221 , 0·651)
95% H.P.D.	(0·161 , 0·651)
99% H.P.D.	(0·073 , 0·651)



Thus, whatever the value of $\mathbf{x}$, the model never associates to $p(\delta_1|\mathbf{x},D)$ a value larger than 0·651, but for specific profiles (high father education, nursery school attended, etc.), this probability may be close to zero. It follows that one may find any profiles among students of private schools, but some profiles are virtually never found in public schools. The relative importance of the factors which influence the type of schools chosen the parents of the students may approximately be read off from their λ-coefficients.

Example 2

A subset of the schools was selected to be used as *experimentals* while the rest would be used as *controls* in testing a particular education programme. It was necessary to

verify that controls and experimentals were well balanced with respect to the attributes considered. Thus, we categorised the data bank into experimentals (d_1) and controls (d_2) and used the methodology described to determine (i) the more relevant classificatory variables (those where bias may be suspected) and (ii) the predictive distribution of the classification probabilities.

The variables sequentially obtained and the corresponding λ-coefficients were:

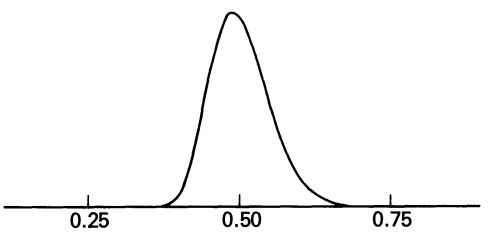
Variables	λ_i
Math. computation ability	1·162
Math. applications ability	−0·535
Orthography ability	0·366
Reading comprehension	−0·357

For each individual we computed the values of the linear discriminant function $t=\lambda' \mathbf{x}$. The Table below summarises the corresponding results within each category group:

	Number Obs.	*Mean*	*Std. Dev.*	*Min.*	*Max.*
Experimental	2173	−0·173	1·014	−3·351	2·614
Control	17 993	0·021	0·998	−11·023	3·446

The predictive distribution of the probability $p(\delta_i|\mathbf{x},D)$ that a random student belongs to an experimental school was found to be

Total Number Obs.	20 166
Mean	0·500
Std. Dev.	0·048
Minimum	0·354
Maximum	1·000
50% H.P.D.	(0·462 , 0·527)
90% H.P.D.	(0·425 , 0·538)
95% H.P.D.	(0·414 , 0·602)
99% H.P.D.	(0·396 , 0·641)



Thus, this probability is practically always within the interval (0·4, 0·6) and is neatly centered in 0·5. Consequently, there is no reason to doubt that experimentals and controls were well balanced with respect to the attributes considered. It is interesting to note, however, that although not very significantly, mathematical computational ability is the poorest balanced among the variables considered, suggesting that, in calculus, controls are marginally better than experimentals.

References

BERNARDO, J.M. (1979a) Expected information as expected utility, *Annals of Statistics*, 7, pp. 686–690.

BERNARDO, J.M. (1979b) Reference posterior distributions for Bayesian inference, *Journal of the Royal Statistics Society*, B-41, pp. 113–147 (with discussion).

BERNARDO, J.M. (1987) Bayesian linear probabilistic classification, in: J. O. BERGER & S. GUPTA (Eds) *Decision Theory and Related Topics IV* (in press).

BERNARDO, J.M. & BERMUDEZ, J.D. (1985) The choice of variables in probabilistic classification, in: J. M. BERNARDO, M. H. DE GROOT, D. V. LINDLEY & A. F. SMITH (Eds) *Bayesian Statistics 2*, pp. 67–81 (with discussion) (Amsterdam, North Holland).

DIACONIS, P. & FREEDMAN, D. (1984) Asymptotics of graphical projection pursuit, *Annals of Statistics*, 12, pp. 793–815.
FISHER, R.A. (1936) The use of multiple measurements in taxonomical problems, *Annals of Eugenetics*, 7, pp. 179–188.
GEISSER, S. (1964) Posterior odds for multivariate normal classification, *Journal of the Royal Statistics Society*, B-26, pp. 69–76.
GOEL, P.K. (1983) Information measures and Bayesian hierarchical models, *Journal of the American Statistical Association*, 78, pp. 408–410.
LINDLEY, D.V. (1982) Scoring rules and the inevitability of probability, *International Statistical Review*, 50, pp. 1–26 (with discussion).
SAVAGE, L.J. (1971) Elicitation of personal probabilities and expectations, *Journal of the American Statistical Association*, 66, pp. 783–801.
TITTERINGTON, D.M. *et al.* (1981) Comparison of discrimination techniques applied to a complex data set of head injured patients, *Journal of the Royal Statistics Society*, A-144, pp. 145–175 (with discussion).