On Predictive Classification of Binary Vectors

Mats Gyllenberg & Timo Koski

Department of Mathematics, University of Turku, Turku, Finland

Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden

The problem of rational classification of a database of binary vectors is analyzed by means of a family of Bayesian predictive distributions on the binary hypercube. The general notion of predictive classification was probably first discussed by S. Geisser. The predictive distributions are expressed in terms of a finite number observables based on a given set of binary vectors (predictors or centroids) representing a system of classes and an entropy-maximizing family of probability distributions. We derive the (non-probabilistic) criterion of maximal predictive classification due to J. Gower (1974) as a special case of a Bayesian predictive classification. The notion of a predictive distribution will be related to stochastic complexity of a set of data with respect to a family of statistical distributions. An application to bacterial identification will be presented using a database of Enterobacteriaceae as in Gyllenberg (1996 c).

A framework for the analysis is provided by a theorem about the merging of opinions due to Blackwell and Dubins (1962). We prove certain results about the asymptotic properties of the predictive learning process.

The work is addressing the following topics of interest:

- automated data analysis
- cluster analysis
- predictive modelling

The maximal predictive classification of Gower (1974) is a method of clustering (unsupervised learning) based on the principle that as many as possible of the properties of the items assigned to the them should be predictable from the class descriptors. In this the statistical techniques of clustering are dismissed since they tend to produce clusters that tell us directly nothing about the members of the classes and thus are potentially irrelevant. In Gower’s discussion of predictivity the characters are binary, here this is in particular motivated by applications to diagnostic microbiology, c.f. Gyllenberg et.al. (1996 c).

From another point of view describing in advance the properties of an item in a certain class, before having inspected the item in detail, is being concerned with uncertainty. Uncertainty or plausibility is representable by probability, as argued by R.T. Cox (1961). The theory of Bayesian classification draws on this thinking, see Cheeseman (1990,1996).

In the sequel we show in particular that Bayesian predictive classification, as introduced
by Geisser (1966) and (1993), in fact generalizes predictiveness in Gower’s sense. Thus probability is indeed involved in the fundamental clustering framework elaborated by Gower.

We consider a given data set $X^t = \{x^{(t)}\}_{t=1}^T$ of $t$ elements of the binary hypercube

$$B^d := \{x \mid x = (x_i)_{i=1}^d, x_i \in \{0, 1\}\}.$$  

By some means $X^t$ has been subdivided into $k$ pairwise disjoint clusters or classes, $c_j = c_j (X^t)$. We refer to the collection of $k$ given classes $\{c_j\}_{j=1}^k$ as a taxonomy. Here and elsewhere in the paper we consider the number of classes $k$ to be given in advance, for a technique determining $k$ from $X^t$ we refer to Gyllenberg et. al. (1996 b).

Let us in addition represent the classes $c_j$ for $j = 1, \ldots, k$ by $a_j = \{a_{1j}, a_{2j}, \ldots, a_{dj}\}$, a binary vector in $B^d$, respectively. At this stage of argument we need not specify how the $a_j$’s are chosen. The idea is that each $x^{(t)}$ in $c_j$ could for some purposes be represented or predicted by the corresponding $a_j$. The error or distortion in this representation can be measured by

$$t_{ij} = \sum_{x^{(t)} \in c_j} |x^{(t)} - a_{ij}|.$$  

We let $t_j$ designate the number of vectors assigned to $c_j$, $j = 1, \ldots, k$. Applying the principle of maximum entropy, c.f. Gyllenberg (1996 a), Bayes formula and a few assumptions of statistical independece between the class-representations we obtain for each class a predictive distribution given by

$$p(z \mid a_j, X^t) = \prod_{i=1}^d \left( \frac{t_{ij} + 1}{t_j + 2} \right)^{|z_i - a_{ij}|} \left( 1 - \left( \frac{t_{ij} + 1}{t_j + 2} \right) \right)^{1-|z_i - a_{ij}|}$$  

for any $z \in B^d$. The distribution $p(z \mid a_j, X^t)$ is a class-conditional probability distribution on $B^d$ that predicts or retrodicts the properties of binary vectors using the knowledge represented by the taxonomy $\{c_j\}_{j=1}^k$. In the notation $p(z \mid a_j, X^t)$ it is tacitly understood that these class-conditional predictive distributions depend on $X^t$ only through those $x^{(t)}$ that are assigned to $c_j$.

The following facts describe the predictive distributions. Let the class $c_j$ for $X^t$ be given and let $t_j = \sum_{x^{(t)} \in c_j} 1$. Let

$$f_{ij} = \frac{1}{t_j} \sum_{x^{(t)} \in c_j} x_i^{(t)}$$  

denote the relative frequency of binary ones in the $i^{th}$ position of $x^{(t)}$’s assigned to $c_j$.

Then the predictor $a_j^* = \{a_{ij}^*, a_{2j}^*, \ldots, a_{dj}^*\}$ defined by

$$a_{ij}^* = \begin{cases} 1 & \text{if } 1/2 < f_{ij} < 1, \\ 0 & \text{if } 0 < f_{ij} < 1/2, \end{cases}$$  

for each $j = 1, \ldots, k$. This is in fact the family of class-conditional predictive distributions.
is the choice of \( a_j \) that maximizes the simultaneous predictive probability of \( c_j \) i.e.

\[
\prod_{x^{(t)} \in c_j} p\left(x^{(t)} \mid a_j, X^t\right) = \prod_{x^{(t)} \in c_j} \prod_{i=1}^{d} \left(\frac{t_{ij} + 1}{t_j + 2}\right)^{i - a_{ij}} \left(1 - \frac{t_{ij} + 1}{t_j + 2}\right)^{1 - i - a_{ij}},
\]

where \( t_{ij} = \sum_{x^{(t)} \in c_j} |x^{(t)}_i - a_{ij}| \). In case there is an \( i \) such that \( f_{ij} = 1/2 \), the binary value of \( a_{ij}^* \) can be chosen arbitrarily.

It holds also that

\[
p\left(a_j^* \mid a_j^*, X^t\right) \geq p\left(z \mid a_j^*, X^t\right)
\]

for every \( z \in B^d \). Let us now suppose that the data base to be used for establishing a taxonomy consists of \( k \) distinct vectors \( A_k = \{a_1, \ldots, a_k\} \) taken as the representers of their respective single member classes. This means that we have \( t_j = 1 \) for all \( j \) and \( t_{ij} = 0 \) for all \( i \) and \( j \).

Let us consider the maximization of the simultaneous predictive probability of \( t \) new strings of \( d \) bits \( x^{(t)}_1, \ldots, x^{(t)}_t \),

\[
\log_2 p\left(x^{(t)}_1, \ldots, x^{(t)}_t \mid a_{j1}, \ldots, a_{jt}, A_k\right) = \sum_{l=1}^{t} \log_2 p\left(x^{(t)}_l \mid a_{jl}, A_k\right)
\]

by attaching each of \( x^{(t)}_1, \ldots, x^{(t)}_t \) to one of the given representers \( a_j \), where \( \log_2 \) is the binary logarithm. In other words we wish to evaluate

\[
\sum_{i=1}^{t} \max_{1 \leq j \leq k} \log_2 p\left(x^{(t)}_l \mid a_j, A_k\right).
\]

But since \( t_j = 1 \) for all \( j \) and \( t_{ij} = 0 \) for all \( i \) and \( j \) we readily obtain

\[
\sum_{i=1}^{t} \max_{1 \leq j \leq k} \log_2 p\left(x^{(t)}_l \mid a_j, A_k\right) = d \cdot t - \sum_{i=1}^{d} \min_{1 \leq j \leq k} |x^{(t)}_i - a_{ij}| - d \cdot t \cdot \log_2 (3).
\]

But this is nothing else but the expression maximized by choice of the codebook \( A_k \) in Gower's work.

Although various posterior predictive distributions are often manipulated texts on pattern recognition, only Ripley (1995) has singled out predictive classification as a distinct topic. The notion of predictiveness is also inherent in the analysis of generalization ability of neural networks, see Bishop (1995), Gyllenberg et.al. (1995).

References:


Cheeseman (1990) P. Cheeseman: On Finding the Most Probable Model. *Computa-


[ Corresponding author:
T. Koski
Department of Mathematics
Royal Institute of Technology
S 100 44 Stockholm
Sweden
e-mail: timo@math.kth.se ]