A classifier based on normalized maximum likelihood model for classes of Boolean regression models

I. Tabus, J. Rissanen and J. Astola

Institute of Signal Processing
Tampere University of Technology
P.O Box 553, SF-33101 Tampere, Finland

ABSTRACT

Boolean regression models are useful tools for various applications in nonlinear filtering, nonlinear prediction, classification and clustering. We discuss here the so-called normalized maximum likelihood (NML) models for these classes of models. Examples of discrimination of cancer types by using the universal NML model for the Boolean regression models indicate its ability to select sets of feature genes discriminating at error rates significantly smaller than those of other discrimination methods.

1 Introduction

We discuss here the NML models for Boolean classes of models. The NML model for the linear regression problem was introduced and analyzed recently, [7]. We restate the classification problem as a modelling problem in terms of a class of parametric models for which the maximum likelihood parameter estimates can be easily computed. We review first the NML model for Bernoulli strings as the solution of a minmax optimization problem. We then introduce a model class for the case when the binary strings to be modelled are observed jointly with several other binary strings (regression variables). We derive the NML model for this model class, provide a fast evaluation procedures and apply it to a classification problem.

The concept of gene expression was introduced four decades ago with the discovery of messenger RNA, when the theory of genetic regulation of protein synthesis was described [5]. The availability of cDNA microarrays makes it possible to measure simultaneously the expressions level for thousands of genes. Gene expression data obtained in microarray experiments may often be discretized as binary or ternary data, the values 1,0,-1 carrying the meanings of overexpressed, normal, and repressed, respectively, which are the needed descriptors when defining regulatory pathways [4].

One possible setting of a classification problem is in terms of a Boolean regression problem. Suppose that the data available is a matrix X, where the entry $x(i,j) \in \{0,1\}$ is a binary (quantized) gene expression, the row index $i \in \{1,\dots,N\}$ identifies the gene, and the column index $j \in \{1,\dots,n\}$ identifies the "patient". We denote by $\underline{x}_j$ the jth column of the matrix X. Furthermore, a class label y_j is known for all patients (e.g. $y_j = 0$ or $y_j = 1$ for the j'th patient having disease type A, or type B, respectively). Our goal is to build Boolean models $\hat{y}_j = f(x_{i_1,j},\dots,x_{i_k,j})$ and to select the set of informative genes, $\{i_1, i_2, \dots, i_k\}$.

2 The NML model for Bernoulli strings

In this section we assume that a Bernoulli variable Y with $P(Y=0)=\theta$ is observed repeatedly n times, generating the string $y^n = y_1,\dots y_n$. We look for a distribution $q(y^n)$ over all strings of length n, such that the ideal codelength $\log\frac{1}{q(y^n)}$ assigned to a particular string y^n by this distribution, is as close as possible to the ideal codelength $\log\frac{1}{P(y^n|\hat{\theta}(y^n))}$ obtainable with the Bernoulli models. In the coding scenario, the decoder is allowed to use a predefined distribution, $q(\cdot)$, but he cannot use the distribution $P(\cdot|\hat{\theta}(y^n))$ because he does not have y^n available. The latter will be the most advantageous distribution in the family $P(y^n|\theta)$ for the string y^n, since it maximizes $P(y^n|\hat{\theta}(y^n))$, and therefore minimizes the ideal codelength $\log\frac{1}{P(y^n|\hat{\theta}(y^n))}$. The distribution $q(y^n)$ is selected such that the "regret" of using $q(y^n)$ instead of $P(y^n|\hat{\theta}(y^n))$, namely,

$$\log\frac{1}{q(y^n)} - \log\frac{1}{P(y^n|\hat{\theta}(y^n))} = \log\frac{P(y^n|\hat{\theta}(y^n))}{q(y^n)}, \tag{1}$$

is minimized for the worst case y^n; i.e.

$$\min_q \max_{y^n} \log\frac{P(y^n|\hat{\theta}(y^n))}{q(y^n)} \tag{2}$$

Theorem 1 *(Shtarkov[9]) The minimizing distribution is given by*

$$q(y^n) = \frac{P(y^n|\hat{\theta}(y^n))}{C_n}, \tag{3}$$

where

$$C_n = \sum_{m=0}^{n} \left(\begin{array}{c} n \\ m \end{array}\right) \left(\frac{m}{n}\right)^m \left(1-\frac{m}{n}\right)^{n-m}. \tag{4}$$

A strong optimality property of the NML models was recently proven in [8], where the following minmax problem was formulated: find the (universal) distribution which minimizes the average regret

$$\min_q \max_g E_g \log \frac{P(Y^n|\hat{\theta}(Y^n))}{q(Y^n)}, \tag{5}$$

where $g(\cdot)$, the generating distribution of the data, and $q(\cdot)$ run through any sets that include the NML model.

Theorem 2 *([8]) The minimizing distribution $q(\cdot)$ in the minmax problem (5) is given by (3) and (4).*

3 The NML model for a Boolean class

We consider a binary random variable Y, which is observed jointly with a binary regressor vector $\underline{X} \in \mathcal{B}^k$. In a useful model class, a carefully selected Boolean function $f : \mathcal{B}^k \to \{0,1\}$ should provide a reasonable prediction $f(\underline{X})$ of Y, in the sense that the absolute error $\mathcal{E} = |Y - f(\underline{X})|$ has a high probability of being 0. Since $\mathcal{E}, Y, f(\underline{X})$ are binary-valued we have $\mathcal{E} = |Y - f(\underline{X})| = Y \oplus f(\underline{X})$, which also implies $Y = f(\underline{X}) \oplus \mathcal{E}$, where $\oplus$ is modulo 2 sum.

We therefore consider a corruption model defined as follows:

$$Y = f(\underline{X}) \oplus \mathcal{E} = \begin{cases} f(\underline{X}) & if \quad \mathcal{E} = 0 \\ \\ \overline{f(\underline{X})} & if \quad \mathcal{E} = 1 \end{cases} \tag{6}$$

where $f(\cdot)$ is a Boolean function and the error $\mathcal{E}$ is independently drawn from a Bernoulli source with parameter θ; i.e., $P(\mathcal{E} = 1) = 1 - \theta$ and $P(\mathcal{E} = 0) = \theta$, or for short

$$P(\mathcal{E} = b) = \theta^{1-b}(1-\theta)^b, \text{ for } b \in \{0,1\} \tag{7}$$

Denote by $\underline{b}_i \in \{0,1\}^k$ the vector having as entries the bits in the binary representation of integer i, i.e., $\underline{b}_0 = [0,\ldots,0,0]$, $\underline{b}_1 = [0,\ldots,0,1]$, etc. Further, define by (6) and (7) the conditional probability for code $\underline{b}_i \in \{0,1\}^k$,

$$P(Y = y|\underline{X} = \underline{b}_i) = \theta^{1-y\oplus f(\underline{b}_i)}(1-\theta)^{y\oplus f(\underline{b}_i)}. \tag{8}$$

The Boolean regression problem will be stated as finding the optimal universal model (in a minmax sense to be specified shortly) for the following class of models:

$$\begin{aligned} &\mathcal{M}(\theta, k, f) = \\ &= \{P(y|f, \underline{b}_i, \theta) = \theta^{(1-y\oplus f(\underline{b}_i))}(1-\theta)^{(y\oplus f(\underline{b}_i))}\} \end{aligned} \tag{9}$$

where $y \in \{0,1\}, \theta \in [0,1], \underline{b}_i \in \{0,1\}^k$.

When the sequence $y^n = y_1 \ldots y_n$ and the sequence of binary regressor vectors $\underline{b}^n = \underline{b}_{i_1}, \ldots, \underline{b}_{i_n}$ are observed, a member of the class $\mathcal{M}(\theta, k, f)$ assigns to the sequence y^n the following probability

$$\begin{aligned} P(y^n|\theta, k, f, \underline{b}^n) &= \prod_{j=1}^n \theta^{(1-y_j\oplus f(\underline{b}_{i_j}))}(1-\theta)^{(y_j\oplus f(\underline{b}_{i_j}))} \\ &= \theta^{n_0}(1-\theta)^{n-n_0}, \end{aligned} \tag{10}$$

where n_0 is the number of zeros in the sequence $\{\varepsilon_j = y_j \oplus f(\underline{b}_{i_j})\}_{j=1}^n$. The ML estimate of the model parameters,

$$(\hat{\theta}(y^n), \hat{f}_{y^n}) = \arg\max_{\theta, f} P(y^n|\theta, k, f, \underline{b}^n), \tag{11}$$

can be obtained in two stages, first by maximizing with respect to f,

$$\max_f P(y^n|\theta, k, f, \underline{b}^n), \tag{12}$$

and observing that the optimal $f(\cdot)$ does not depend on θ. For a fixed $\theta > 0.5$, the function $P(y^n|\theta, k, f, \underline{b}^n) = \theta^{n_0}(1-\theta)^{n-n_0}$ decreases monotonically with n_0, and (12) is maximized by maximizing n_0, or, equivalently, by minimizing $n - n_0$

$$\begin{aligned} \min_f (n - n_0) &= \min_f \sum_{j=1}^n |y_j - f(\underline{b}_{i_j})| \\ &= \min_f \sum_{\ell=0}^{2^n} m_{\ell_0} f(\underline{b}_\ell) + m_{\ell_1}(1 - f(\underline{b}_\ell)). \end{aligned} \tag{13}$$

Equation (13) shows that f should be optimal for the mean absolute error (MAE) criterion. It can also be seen that the assignment of $f(\underline{b}_\ell)$ depends solely on m_{ℓ_0}, m_{ℓ_1}, and the solution is

$$\hat{f}_{y^n}(\underline{b}_\ell) = \begin{cases} 0 & if \quad m_{\ell_0} \geq m_{\ell_1} \\ 1 & if \quad m_{\ell_0} < m_{\ell_1} \end{cases}, \tag{14}$$

which can be readily computed from the data set. Denote by $n_0^*(y^n)$ the number of zeros in the sequence $\{\varepsilon_j = y_j \oplus \hat{f}_{y^n}(\underline{b}_{i_j})\}_{j=1}^n$. To completely solve the ML estimation problem we have to find

$$\max_\theta P(y^n|\theta, k, \hat{f}_{y^n}, \underline{b}^n), \tag{15}$$

for which the maximizing parameter is $\hat{\theta}(y^n) = \frac{n_0^*(y^n)}{n}$. Therefore

$$\begin{aligned} &P(y^n|\hat{\theta}(y^n), k, \hat{f}_{y^n}, \underline{b}^n) \\ &\quad = \left(\frac{n_0^*(y^n)}{n}\right)^{n_0^*(y^n)} \left(1 - \frac{n_0^*(y^n)}{n}\right)^{n-n_0^*(y^n)}. \end{aligned} \tag{16}$$

We need to define a distribution $q(y^n)$ over all possible sequences y^n, which is the best in the minmax sense

$$\min_q \max_{y^n} \frac{P(y^n|\hat{\theta}(y^n), k, \hat{f}_{y^n}, \underline{b}^n)}{q(y^n)}, \tag{17}$$

which is clearly given by the NML model,

$$q(y^n) = \frac{P(y^n|\hat{\theta}(y^n), k, \hat{f}_{y^n}, \underline{b}^n)}{C_n(k, \underline{b}^n)}, \tag{18}$$

where

$$\begin{aligned} &C_n(k, \underline{b}^n) = \\ &= \sum_{y^n} \left(\frac{n_0^*(y^n)}{n}\right)^{n_0^*(y^n)} \left(1 - \frac{n_0^*(y^n)}{n}\right)^{n-n_0^*(y^n)} \end{aligned} \tag{19}$$

We remark that n_0^* depends on y^n through $\hat{f}_{y^n}$ in a complicated manner. When $k = 0$, the normalization factor is $C_n(0, \underline{b}^n) = C_n$, given in (4).

Alternative expressions for the coefficient $C_n(k, \underline{b}^n)$ provide faster evaluation. We need to specify the distinct elements in the set $\{\underline{b}_\ell | \underline{b}_\ell \in \underline{b}^n\}$ as $\{\underline{b}_{j_1}, \dots, \underline{b}_{j_K}\}$, and denote by z^q the subsequence of y^n observed when the regressor vector is $\underline{b}_{j_q}$. Let n_q be the length of the subsequence z^q having m_q zeros.

We observe that (19) can be alternatively expressed as

$$C_n(k, \underline{b}^n) = \sum_{n_1^*=0}^{n} \left(\frac{n_1^*}{n}\right)^{n_1^*} \left(1 - \frac{n_1^*}{n}\right)^{n-n_1^*} S_{K,n_1,\dots,n_K}(n_1^*),$$

where $S_{K,n_1,\dots,n_K}(n_1^*)$ is the number of sequences y^n having $n_1^* = \sum_{q=1}^K \min(m_q, n_q - m_q)$ ones in the residual sequence. The numbers $S_{K,n_1,\dots,n_K}(n_0^*)$ can be easily computed, recursively in K. Denote first

$$h_\ell(m) = \begin{cases} 0 & if \quad m > \frac{n_\ell}{2} \\ \begin{pmatrix} n_\ell \\ m \end{pmatrix} & if \quad m = \frac{n_\ell}{2} \\ 2\begin{pmatrix} n_\ell \\ m \end{pmatrix} & else \end{cases}, \quad (20)$$

which is the number of sequences of n_ℓ bits, having either m bits set to 1, or $n_\ell - m$ bits set to 1, for $0 \leq m \leq \frac{n_\ell}{2}$. By combining each of the $S_{K-1,n_1,\dots,n_{K-1}}(n_1^* - m_K)$ sequences having $n_1^* - m_K$ ones in the residual sequence, with each of the $h_K(m_K)$ sequences having either m_K bits set to 1, or $n_K - m_K$ bits set to 1, we get sequences having $(n_1^* - m_K) + \min(m_K, n_K - m_K) = n_1^*$ bits of 1 in their residual sequence. Therefore the following recurrence relation holds:

$$S_{K,n_1,\dots,n_K}(n_1^*) = \sum_{m_K=0}^{n_K} h_K(m_K) S_{K-1,n_1,\dots,n_{K-1}}(n_1^* - m_K), \quad (21)$$

where, by convention, $S_{K-1,n_1,\dots,n_{K-1}}(n_1^* - m_K) = 0$ for negative arguments, $n_1^* - m_K < 0$.

We note that the recurrence is simply a convolution sum, $S_{K,n_1,\dots,n_K} = h_K \otimes S_{K-1,n_1,\dots,n_{K-1}}$, and from here we conclude that

$$S_{K,n_1,\dots,n_K} = h_1 \otimes h_2 \otimes \dots \otimes h_K. \quad (22)$$

We can easily see that $S_{K_1,n_1,\dots,n_{K_1}}(i) = 0$ for $i > \frac{\sum_q^{K_1} n_q}{2}$, due to the fact that the optimal residual sequence cannot have more than $\frac{\sum_q^{K_1} n_q}{2}$ ones. Also, from (20) we note that only $\frac{1}{2^K} \prod_{q=1}^K n_q$ terms have to be added when evaluating all convolution sums (22).

4 Experimental results

We illustrate the classification based on NML model for classes of Boolean regression models using the microarray DNA data Leukemia (ALL/AML) of [3], publicly available at http://www-genome.wi.mit.edu/MPR/. The microarray contains 6817 human genes, sampled from 72 cases of cancer, of which 47 are of ALL type and 25 of AML type. The data is preprocessed as recommended in [3] and [2]. The resulting data matrix $\tilde{X}$ has 3571 rows and 72 columns.

We design a two level quantizer by using the LBG algorithm [6] and the decision threshold results at 2.6455. All the entries in the matrix $\tilde{X}$ are used as a training set (but we note that no information about the true classes is used during the quantization stage). The entries in the matrix $\tilde{X}$ are quantized to binary values, resulting in the binary matrix X.

4.1 Extending the classification for unseen cases of the Boolean regressors

The Boolean regressors observed in the training set may not span over all 2^k possible binary vectors. If the binary vector $\underline{b}_q$ is not observed in the training set, the classification decision $f^*(\underline{b}_q)$ remains undecided during the training stage. We decide the value of $f^*(\underline{b}_q)$ by using nearest neighbor voting, taking as decision the majority vote of the neighbors $\underline{b}_\ell$ situated at Hamming distance 1, for which $f^*(\underline{b}_\ell)$ was decided during the training stage. If after voting there is still tie we take the majority vote of the neighbors at Hamming distance 2, and continue if necessary until a clear decision is reached.

4.2 Estimation of classification errors achieved with Boolean regression models with $k = 3$

The Leukemia data set was considered recently in a study comparing several classification methods[2]. The evaluation of the performance is based there on the classification error estimated in a crossvalidation 2:1 experiment. In order to compare our classification results with the results in [2], we estimate the classification error in the same way, namely dividing at random the 72 patient set into a training set of $n_T = 48$ patients and a test set of $n_s = 24$ patients, find the optimal predictor $f^*(\cdot)$ over the training set, classify the test set by using the predictor $f^*(\cdot)$ (the extension for cases unseen in the training set is done as in Section 4.1), and count the number of classification errors produced over the test set. The random split is repeated a number of $n_r = 10000$ times, and the estimated classification error is computed as the percentage of the total number of errors observed in the $(n_r \cdot n_s)$ test classifications. For comparison, we mention that the best classification methods tested in [2] have classification errors higher than 1%. As we can observe in Table 1 there are several predictors with three genes, achieving classification rates as low as 0.004%. We note a remarkable consensus in ranking of the gene triplets, according to the NML codelength and to the estimated classification error rates.

As for the genes involved in the optimal predictors of Table 1, we note that five genes belong to the

Table 1: The best 18 triplets of genes for predicting the class label according to the NML model for the class $\mathcal{M}(\theta, 3, f)$.

Codelength	Classification error [%]	Triplet of Genes			Gene accession numbers		
6.9	0.912	1834	2288	5714	M23197	M84526	HG1496-HT1496
7.9	**0.010**	**1834**	**3631**	**6277**	**M23197**	**U70063**	**M30703**
7.9	0.891	758	4250	4342	D88270	X53586	X59871
8.0	0.652	2288	4847	6376	M84526	X95735	M83652
8.7	**0.008**	**1834**	**3631**	**5373**	**M23197**	**U70063**	**S76638**
8.7	**0.007**	**1834**	**3631**	**6279**	**M23197**	**U70063**	**X97748**
8.7	0.910	1144	1217	1882	J05243	L06132	M27891
8.8	0.649	302	2288	6376	D25328	M84526	M83652
8.8	0.055	1144	1834	1882	J05243	M23197	M27891
8.8	0.063	1834	1882	6049	M23197	M27891	U89922
8.8	**0.004**	**1144**	**1882**	**5808**	**J05243**	**M27891**	**HG2981-HT3127**
8.8	0.584	2288	3932	6376	M84526	U90549	M83652
8.9	0.558	2288	5518	6376	M84526	X95808	M83652
8.9	0.560	1399	2288	6376	L21936	M84526	M83652
8.9	0.620	1241	2288	6376	L07758	M84526	M83652
8.9	0.605	2288	3660	6376	M84526	U72342	M83652
8.9	0.582	2288	4399	6376	M84526	X63753	M83652
8.9	0.556	2288	4424	6376	M84526	X65867	M83652

set of 50 “informative” genes selected in [3], namely $M23197, M84526, M27891, M83652, X95735$.

5 Conclusion

Boolean regression classes of models are powerful modelling tools having associated NML models which can be easily computed and used in MDL inference, in particular for factor selection.

The use of MDL for classification by resorting to the class of Boolean models provides a principled and effective classification method, as we exemplify with the important application of cancer classification based on gene expression data. The NML model for the class $\mathcal{M}(\theta, k, f)$ was used for the selection of informative feature genes. When using the sets of feature genes selected by NML model, we achieved classification error rates significantly lower than those reported recently for the same data set.

References

[1] A. Barron, J. Rissanen, Y. Bin. The minimum description length principle in coding and modeling. IEEE Trans. on Information Theory, Special commemorative issue: Information Theory 1948-1998, vol.44, no. 6, 2743–2760, Oct. 1998.

[2] S. Dudoit, J. Fridlyand, T. P. Speed. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Dept. of Statistics University of California, Berkeley, Technical Report 576, 2000.

[3] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, Vol. 286, pp. 531-537, Oct. 1999.

[4] S. Kim, E.R. Dougherty. Coefficient of determination in nonlinear signal processing. Signal Processing, 80:2219–2235, 2000.

[5] F. Jacob, J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. Journal of Molecular Biology, Vol. 3, 318-356, 1961.

[6] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantization design. *IEEE Transactions on Communications*, 28:84–95, January 1980.

[7] J. Rissanen. MDL Denoising. IEEE Trans. on Information Theory, vol. IT-46, No. 7, 2537–2543, Nov. 2000.

[8] J. Rissanen. Strong optimality of the normalized ML models as universal codes and information in data. IEEE Trans. on Information Theory, vol.IT-47, No. 5, 1712–1717, July 2001.

[9] Yu.M. Shtarkov. Universal sequential coding of single messages. Translated from Problems of Information Transmission, Vol. 23, No. 3, 3–17, July-September 1987.

[10] Ioan Tabus and Jaakko Astola. On the Use of MDL Principle in Gene Expression Prediction. Journal of Applied Signal Processing, Volume 2001, No. 4, December 2001.