Bayesian Learning of unobservable output
aggregating multiple weak classifiers

Badih Ghattas
Universit de la Mediterrane,
CNRS, Luminy, Marseille, France.
E-mail: ghattas@univmed.fr

Gonzalo Perera
Laboratorio de Probabilidad y Estadística,
Facultad de Ingeniería, Universidad de la República, Herrera y Reissig 565.
Montevideo, Uruguay.
E-mail: gperera@fing.edu.uy

Juan Piccini*
Laboratorio de Probabilidad y Estadística,
Facultad de Ingeniería, Universidad de la República, Herrera y Reissig 565.
Montevideo, Uruguay.
E-mail: piccini@fing.edu.uy
*Corresponding author

Abstract: In this paper we propose a bayesian classification rule in a context where the training sample is not complete. The input $X$ is observable, but the output $Y$ is not, and we only utilize the result of weak classifiers instead of $Y$. Our method is motivated by epidemiological and bioinformatics problems. We prove the consistency of our method and illustrate its efficiency using simulations. Although up to our knowledge there are no similar algorithms for unobservable output, we compared in our simulations to supervised approaches.

Keywords: Bayesian Learning, Boosting, Unobservable Output, MAP.

Biographical notes:
X
X

1 Introduction

In some situations where we plan to learn the relation between an output $Y$ using a vector of explanatory variables $X$, the output $Y$ is non observable but some
weak classifiers may be available. This is the case for example in epidemiological studies where we wish to decide whether patients are affected or not affected, the classification being based on a clinical test. Such a test, when done again by the same person or another one, might come out differently. So the right status of the patient may be seen as a non observed random variable \( Y \), and the outcome of the test may be considered as another random variable \( T \) or a weak classifier.

Another example encountered in Bioinformatics and which inspired our work, is the one concerned with learning whether a DNA sequence is or is not of particular interest (PAS, Poly-Adenylation Sites). The outcome of each sequence is unobservable, but it may be estimated using some limited numerical or statistical models.

In many cases the fact that the outcome is unobservable may be due to the cost of it’s observation and/or the very high number of observations where it should be measured.

In the sequel we focus on an epidemiological example, however the same scheme can be applied to many other areas. Our approach is close in spirit to Boosting, since we perform an efficient classifier based on a set of weak classifiers. One fundamental difference is that, unlike boosting, the weak classifiers that we call "tests" are not learned using a training data \((X,Y)\) but are either given by experts or estimated only using \(X\).

For sake of clarity and without loss of generality, we consider only the case of binary classification, thus \( Y \in \{0, 1\} \). We wish to estimate the probability for an observation \( x \) to belong to class 1. Our classifier is obtained by using a sequence of independent tests \( T^j, j = 1, 2, \ldots, k \) made over each observation \( x \), and by updating the maximum a posterior probability (MAP) for \( x \) to be in class 1. The update of these probabilities is done as follows: it increases whenever a new test \( T^j \) assigns class 1 to an observation and decreases in the opposite case. We prove the efficiency of our algorithm, and give a method for computing the exact number of iterations required to achieve a fixed performance.

The outline of our paper is as follows: Section two introduces our algorithm for the construction of the MAP classifier. Section three gives some proofs for the consistence of our method and section four illustrates our procedure with a simulation study. Although up to our knowledge no algorithms may be compared directly to our, we did some comparisons using simulated models as \( Y \) is also simulated.

2 Our MAP classifier

2.1 Notations and definitions

Assume that an individual \( w \) is chosen from a given population \( \Omega \) and that we denote by \( X(w) \) a vector containing a set of explanatory variables for \( w \) (age, weight, gender, etc.), taking values on a finite set \( S_X \). Denote by \( Y(w) \) his status with respect to a given pathology: \( Y(w) = 1 \) means that \( w \) is affected, \( Y(w) = 0 \) otherwise. Our main problem is that this status \( Y \) is non observable.
Bayesian Learning of unobservable output

Assume a clinical test is available for the detection of the pathology:

\[ T : S_X \times \{0,1\} \rightarrow \{0,1\} \]

where \( T(X(w),Y(w)) = 1 \) when the test indicates that \( w \) is affected by the pathology and \( T(X(w),Y(w)) = 0 \) otherwise.

A measurable structure is assumed over \( S_X \times \{0,1\} \), thus \( T \) is taken to be a measurable map, equivalently a random variable on \( \mathcal{C} = \{ f : S_X \times \{0,1\} \rightarrow \{0,1\} \} \). As usual to simplify the notations we will omit \( w \).

A probability \( P \) is considered on the underlying measurable space. The test \( T \) has two types of misclassification errors: false positives and false negatives, with given probabilities for fixed values of \( X \).

More precisely, we consider \( \alpha : S_X \rightarrow [0,1], \beta : S_X \rightarrow [0,1] \) measurable maps, such that:

\[
\forall x \in S_X, \, \alpha(x) = P(T(x,0) = 1), \beta(x) = P(T(x,1) = 0)
\]

In addition, we assume that \( T \) behaves as a weak learner (see ?), that is

\[
\forall x \in S_X, \, 0 < \alpha(x) < 1/2, \, 0 < \beta(x) < 1/2
\]

Consider now \( n \) individuals, selected at random:

\[(X_1,Y_1),\ldots,(X_n,Y_n) \ iid \sim \rho\]

where \( \rho \) is the joint distribution of \((X,Y)\). Denote \( \pi \) the marginal distribution of \( X \) and \( p(\cdot|x) \) the conditional distribution of \( Y \) given \( X = x \), i.e.,

\[
\forall x \in S_X, \, y \in \{0,1\}, \, \pi(x) = P(X = x), \, p(y|x) = P(Y = y|X = x).
\]

We will assume that

\[
\forall x \in S_X, \, \pi(x) > 0.
\]

2.2 Constructing the bayesian classifier

Recall that our main problem is that the “Y’s” are not observable and we do not know if \( w \) is affected or not. What we can observe is the result of the clinical test for each one of the individuals, repeated several times. Therefore, assume \((T_1^{(j)})_{1 \leq i \leq n, j=1,2,\ldots} \) are iid. At the \( j^{th} \) repetition of the test we observe

\[(X_1, T_1^{(j)}(X_1,Y_1)),\ldots,(X_n, T_n^{(j)}(X_n,Y_n)).\]

Since \( T_1^{(j)},\ldots,T_n^{(j)} \) are random functions on \( S_X \times \{0,1\} \), two individuals \( X_i, X_l \) such that \( X_i = X_l = x \), may have different test results, even in the case where \( Y_i = Y_l = y \), since \( T_i^{(j)}(x,y) \neq T_l^{(j)}(x,y) \) is perfectly possible. Observe also that for the same individual, two replicas of the test may give different results, since \( T^{(j)}(X_i,Y_i) \neq T^{(l)}(X_i,Y_i) \) for \( j \neq l \) is possible.

We perform \( k \) times, \( k \geq 3 \) the test to all the observations. Hence at the \( k^{th} \) iteration a sample \((X_1, T_1^{(k)}(X_1,Y_1)),\ldots,(X_n, T_n^{(k)}(X_n,Y_n))\) is available. Set
\[ Z_i^{(k)} = \sum_{j=1}^{k} T^{(j)}(X_i, Y_i). \] For fixed values \( X = x \) and \( Y = y \), \((T^{(j)}(X,Y))_{j=1,2,...,k}\) are random Bernoulli independent variables and \( Z_i^{(k)} \) follows a Binomial distribution, denoted as usual Bin\((k,p)\), whose parameters depend on the value of \((X,Y)\).

Using (1) we have:

\[ U = (Z_i^{(k)}|Y_i = 1, X_i = x) \sim \text{Bin}(k, 1 - \beta(x)) \]
\[ V = (Z_i^{(k)}|Y_i = 0, X_i = x) \sim \text{Bin}(k, \alpha(x)) \] (3)

We apply now Bayes rule for \( P(Y_i = 1|X_i = x, Z_i^{(k)} = z) \) setting \( \eta \in \[0, 1\] \) and \( z \in 1, 2, \ldots, k \). We have: \( P(Y_i = 1|X_i = x, Z_i^{(k)} = z) > \eta \) if and only if (iff)

\[
P(U = z) p(1|x) \pi(x) \over P(U = z) p(1|x) \pi(x) + P(V = z) p(0|x) \pi(x) > \eta
\]

which is equivalent also to:

\[
P(U = z) p(1|x) \pi(x) \over P(V = z) p(0|x) \pi(x) > \frac{\eta}{1 - \eta}.
\]

Therefore our MAP classifier is defined by:

\[
\hat{Y}_i^{(k)} = \begin{cases} 1 \text{ if } & P(U = z) p(1|x) \over P(V = z) p(0|x) \geq \left( \frac{\eta}{1 - \eta} \right) \\ 0 \text{ otherwise} & 
\end{cases}
\] (4)

Table ?? summarizes our algorithm.

| For each observation \((x,y)\), estimate \(\alpha(x), \beta(x), \pi(x), p(1|x)\) |
| Set \(u_0^i = 1\) |
| For \(j = 1, 2, \ldots, k\) |
| For each observation \(i\) in the sample : |
| set \(T_i^j(X_i, Y_i)\) |
| Make the update: \(u_i^j = u_i^{j-1} \left( \frac{1 - \beta}{\alpha} \right)^{T_i^j} \left( \frac{\beta}{1 - \alpha} \right)^{1 - T_i^j} \) |
| Set \(\hat{Y}_i^j = 1\) if \(u_i^j \left( \frac{p(1|x)}{p(0|x)} \right) > \frac{\eta}{1 - \eta} \) |

Table 1  Bayesian learning algorithm with non observable output.

Our classifier depends on the parameters \(\alpha(x), \beta(x), \pi(x)\) and \(p(y|x)\). These parameters are -in general- unknown, but we will show that we are able to estimate them by solving a nonlinear system of equations.
In this section we will first describe how we estimate the unknown parameters $\alpha(x)$, $\beta(x)$, $\pi(x)$ and $p(1|x)$, and then prove that these estimates are consistent.

Let us first fix some notation:

\[
\begin{align*}
\pi_n(x) &= \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i = x\}} \\
N_n^{(k)}(x) &= \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i = x, \sum_{j=1}^{k} T^{(i)}(X_i, Y_i) = k\}}
\end{align*}
\]

(5)

$\pi_n(x)$ is the proportion of times where $x$ is observed in the data sample, while $N_n^{(k)}$ is the proportion of times where $x$ is observed, the $k$ tests being positive.

Observe that $N_n^{(k)} \leq \pi_n(x)$ and

\[
\lim_n N_n^{(k)}(x) = p(0|x)\pi(x)\alpha(x)^k + p(1|x)\pi(x)(1 - \beta(x))^k.
\]

Thus,

\[
\lim_n N_n^{(k)}(x) = p(0|x)\pi(x)\alpha(x)^k + p(1|x)\pi(x)(1 - \beta(x))^k.
\]

By the Strong Law of Large Numbers, as $n \to \infty$, we have

\[
\lim_n \pi_n(x) = P(X_1 = x) = \pi(x)
\]

and

\[
\lim_n \frac{N_n^{(k)}(x)}{\pi_n(x)} = p(0|x)\alpha(x)^k + p(1|x)(1 - \beta(x))^k
\]

(6)

Then, using (??) for $k = 1, 2, 3$, setting

\[
v_n(x) = \begin{pmatrix} v_n^{(1)}(x) \\ v_n^{(2)}(x) \\ v_n^{(3)}(x) \end{pmatrix} = \begin{pmatrix} N_n^{(1)}(x)/\pi_n(x) \\ N_n^{(2)}(x)/\pi_n(x) \\ N_n^{(3)}(x)/\pi_n(x) \end{pmatrix}, \quad \theta(x) = (p(0|x), \alpha(x), 1 - \beta(x))
\]

we get a non linear system

\[
\forall x \in S_X, \ G(\theta(x)) = \lim_n v_n(x) = v(x)
\]

(7)
where
\[ G(p,a,b) = \begin{pmatrix} pa + (1-p)b \\ pa^2 + (1-p)b^2 \\ pa^3 + (1-p)b^3 \end{pmatrix} \]  

(8)

with the following constraints: 0 < \( v_n^{(3)}(x) < v_n^{(2)}(x) < v_n^{(1)}(x) < 1 \), \( p \in [0,1], 0 < a < 1/2 < b < 1 \).

Thus we can estimate \( \theta(x) \) by \( \theta_n(x) = (p_n(0|x), \alpha_n(x), 1 - \beta_n(x)) \) defined as the solution of the constrained non linear system:

\[ \forall x \in S_X, G(\theta_n(x)) = v_n(x) \]  

(9)

The consistency of the estimators comes from the regularity of \( G \). The solution may be obtained by minimizing the least squares error between the left and right hand sides of the system. We may augment the system by adding equations corresponding to other values of \( k \) in (??), this should accelerate the convergence of the solutions.

Let us denote

\[ M(p,a,b) = \frac{1}{(a-b)^2} \begin{pmatrix} \frac{6a}{a-b} & \frac{3(a+b)}{a-b} & -2 \\ \frac{2(a+b)b}{p} & \frac{a+2b}{p} & \frac{1}{p} \\ \frac{(a+2b)a}{(1-p)} & \frac{2a+b}{(1-p)} & \frac{1}{(1-p)} \end{pmatrix} \]

and \( \Sigma(x) \) the matrix whose general term is:

\[ (\Sigma(x))_{s,t} = \frac{1}{\pi(x)}[\pi^{(s\vee t)}(x) - \pi(x)\pi^{(s)}(x)\pi^{(t)}(x)] \]

where \( s \vee t = \max(s,t) \).

**Theorem 3.1:**

Using the same notations as above, one has:

1. For any \( x \in S_X \), \( \theta_n(x) \) is a consistent estimator of \( \theta(x) \).

2. \( \sqrt{n}(\theta_n(x) - \theta(x)) \)

   converges in distribution to \((u(x), v(x), r(x))\), a centered random Gaussian three-dimensional vector with covariance matrix:

   \[ M(\theta(x))\Sigma(x)M(\theta(x))^T \]

3. For any \( x \in S_X \), \( M(\theta_n(x))\Sigma_n(x)M(\theta_n(x))^T \) is a consistent estimator of \( M(\theta(x))\Sigma(x)M(\theta(x))^T \), with

   \[ (\Sigma_n(x))_{s,t} = \frac{1}{\pi_n(x)}[\pi_n^{(s\vee t)}(x) - \pi_n(x)\pi_n^{(s)}(x)\pi_n^{(t)}(x)] \]
Bayesian Learning of unobservable output

It is straightforward to check that the Jacobian matrix of \( G \) is

\[
J_G(p, a, b) = \begin{pmatrix}
a - b & p & 1 - p \\
a^2 - b^2 & 2pa & 2(1 - p)b \\
a^3 - b^3 & 3pa^2 & 3(1 - p)b^2
\end{pmatrix}
\]

whose determinant is

\[
\det J_G(p, a, b) = p(a - b)^4(1 - p) > 0
\]

Therefore, by the inverse function theorem, \( G \) admits a \( C^\infty \) local inverse function whose Jacobian matrix is

\[
J_{G^{-1}}(v_1, v_2, v_3) = J_G(p, a, b)^{-1}
\]

with

\[
(p, a, b) = G^{-1}(v_1, v_2, v_3).
\]

Since

\[
\theta_n(x) = G^{-1}(v_n(x)), \quad \theta(x) = G^{-1}(v(x)),
\]

\[
\lim_{n} v_n(x) = v(x) \text{ a.s.}
\]

and \( G^{-1} \) is continuous, i) follows.

ii) It is very easy to check that

\[
\sqrt{n} (v_n(x) - v(x))
\]

converges in distribution to a centered, Gaussian random variable with covariance matrix \( \Sigma(x) \). We will first compute the general term \( \Sigma_{m,i}(x) \). For this computation we simplify the notations using the following:

\[
N^{(k)}_n = N^{(k)}_n(x) = \frac{1}{n} \sum_{i=1}^n W^{(k)}_i where W^{(k)}_i = 1_{\{X_i = x, \sum_{j=1}^n T^{(j)}(X_i, Y_i) = k\}}
\]

Recall that

\[
E(N^{(k)}_n) = E(W^{(k)}_1) = P\{X_i = x, \sum_{j=1}^n T^{(j)}(X_i, Y_i) = k\} = \pi(x)v^{(k)}(x)
\]

and

\[
\forall i, \quad W^{(k)}_i \sim Bernoulli(\pi(x)v^{(k)}(x))
\]

We first compute the covariance of \( N^{(k)}_n \), \( k = 1..3 \).

\[
\text{Cov}(N^{(s)}_n, N^{(t)}_n) = E(N^{(s)}_n N^{(t)}_n) - E(N^{(s)}_n)E(N^{(t)}_n) = E(N^{(s)}_n N^{(t)}_n) - \pi^2(x)v^{(s)}(x)v^{(t)}(x)
\]
The first term of the right side may be written as follows:

\[
E \left( N^{(s)}_n N^{(t)}_n \right) = \frac{1}{n^2} E \left( \sum_{i=1}^{n} W^{(s)}_i \sum_{i=1}^{n} W^{(t)}_i \right) \\
= \frac{1}{n^2} E \left( \sum_{i=1}^{n} W^{(s)}_i W^{(t)}_i + \sum_{i=1}^{n} W^{(s)}_i \left( \sum_{j=1, j \neq i}^{n} W^{(t)}_j \right) \right) \\
= \frac{1}{n^2} \sum_{i=1}^{n} E(W^{(s)}_i W^{(t)}_i) + \frac{1}{n} E \left( W^{(s)}_i \left( \sum_{j=1, j \neq i}^{n} W^{(t)}_j \right) \right) \\
= \frac{1}{n} E \left( W^{(s)}_1 W^{(t)}_1 \right) + \frac{1}{n} E \left( W^{(s)}_1 \right) (n-1) E \left( W^{(t)}_1 \right) \\
= \frac{1}{n} E \left( W^{(s)}_1 W^{(t)}_1 \right) + \frac{n-1}{n} E \left( W^{(s)}_1 \right) E \left( W^{(t)}_1 \right) \\
= \frac{1}{n} E \left( W^{(s)}_1 W^{(t)}_1 \right) + \frac{n-1}{n} \pi^2(x)v^{(s)}(x)v^{(t)}(x) \\
\tag{11}
\]

The random variable \( W^{(s)}_1 W^{(t)}_1 \) is binary and equals 1 when the factor of higher
index among \( s \) and \( t \) equals one. Thus,

\[
W^{(s)}_1 W^{(t)}_1 = 1 \text{ if and only if } \sum_{j=1}^{s \vee t} T^{(j)}(X_j, Y_j) = s \vee t 
\]

so

\[
E \left( W^{(s)}_1 W^{(t)}_1 \right) = P(W^{(s)}_1 W^{(t)}_1 = 1) = \pi(x)v^{(s \vee t)}(x) 
\]

Plugging this in (13) and aggregating with (12) we have for the covariance

\[
\text{Cov}(N^{(s)}_n, N^{(t)}_n) = \frac{\pi(x)}{n} v^{(s \vee t)}(x) + \frac{n-1}{n} \pi^2(x)v^{(s)}(x)v^{(t)}(x) - \pi^2(x)v^{(s)}(x)v^{(t)}(x) \\
= \frac{\pi(x)}{n} [v^{(s \vee t)}(x) - \pi(x)v^{(s)}(x)v^{(t)}(x)] \\
= \frac{\pi(x)}{n} [v^{(s \vee t)}(x) - \pi(x)v^{(s)}(x)v^{(t)}(x)] \\
\tag{12}
\]

Finally,

\[
\text{Cov} \left( \sqrt{n} \frac{N^{(s)}_n}{\pi_n(x)}, \sqrt{n} \frac{N^{(t)}_n}{\pi_n(x)} \right) = \frac{n}{\pi^2_n(x)} \pi(x) \text{Cov}(N^{(s)}_n, N^{(t)}_n) \\
= \frac{n}{\pi^2_n(x)} \pi(x) \left[ v^{(s \vee t)}(x) - \pi(x)v^{(s)}(x)v^{(t)}(x) \right] \\
\rightarrow_{n \rightarrow \infty} \frac{1}{\pi(x)} \left[ v^{(s \vee t)}(x) - \pi(x)v^{(s)}(x)v^{(t)}(x) \right] \\
\tag{13}
\]

We now need the following property: If a sequence of \( R^d \)-valued random
variables \( Z_n \) is such that for a fixed \( \mu \in R^d \), \( \sqrt{n} (Z_n - \mu) \) converges in distribution
to a centered Gaussian random variable with covariance matrix \( \Sigma \), and \( H : R^d \rightarrow R^d \) is of class \( C^2 \) in a neighborhood of \( \mu \), then

\[
\sqrt{n} (H(Z_n) - H(\mu)) 
\]

converges in distribution to a centered Gaussian random variable with covariance matrix

\[
J_H(\mu)\Sigma J_H(\mu)^T. 
\]
Bayesian Learning of unobservable output

Using some elementary algebra it is easy to check that:

\[ J_{G^{-1}}(p, a, b) = J_G(p, a, b)^{-1} = \frac{1}{(a - b)^2} \begin{pmatrix} \frac{-6a}{a-b} & \frac{3(a+b)}{a-b} & \frac{2}{a-b} \\ \frac{(2a+b)b}{p} & \frac{a+2b}{p} & \frac{1}{p} \\ \frac{(a+2b)(a)}{p(1-p)} & \frac{2a+b}{p(1-p)} & \frac{1}{p(1-p)} \end{pmatrix} = M(p, a, b) \]

Taking \( H = G^{-1} \) and using \( \theta_n(x) = H(v_n(x)) \) and \( \theta(x) = H(v(x)) \), ii) follows.

iii) As \( M(\theta_n(x)) \Sigma_n(x) M(\theta_n(x))^T \) is a continuous function of \( \theta_n \) for all \( x \) and \( \theta_n \) converges a.s. to \( \theta \) we get the result. \( \square \)

4 Consistency of the MAP classifier

In this section we will prove the consistency of our estimator defined in ??, and give a lower bound \( K \) for the number of tests necessary to achieve a desired precision.

Let us now study the behavior of our classification rule as \( k \to \infty \).

We introduce the following definitions. Set

\[ C(p, q, \tau) = \frac{\log \tau}{\log \left( \frac{1-p}{p} \frac{1-q}{q} \right)} \quad A(p, q) = \frac{\log \left( \frac{p}{1-q} \right)}{\log \left( \frac{1-p}{p} \frac{1-q}{q} \right)} + (1-p) \]

and let

\[ a_{m,M} = \min_{p,q \in [m,M]} A(p, q) \]

\[ C_{m,M}(\tau) = \max_{p,q \in [m,M]} C(p, q, \tau) \]

Finally for \( \varepsilon \in [0,1] \), define

\[ k_{m,M}(\varepsilon, \tau) = \min \left\{ k \in \mathbb{N} : k > \frac{2C_{m,M}(\tau)}{a_{m,M}}, k > \log(\frac{1}{\varepsilon}) \left( \frac{6 + 4a_{m,M}}{3a_{m,M}} \right) \right\}. \]

The following elementary Lemma, concerning the Binomial distribution, plays an essential role in the sequel.

**Lemma 4.1:** Let \( k \in \mathbb{N}, p, q \in [m,M], 0 < m < M < 1/2, \) \( U, V \) real random variables such that

\[ U \sim \text{Bin}(k, 1-p), V \sim \text{Bin}(k, q), \]

and \( \tau > 0 \).

Define

\[ g_k(z) = \frac{P(U = z)}{P(V = z)} \]
1. We have: $m, M > 0$, and for a fixed $\tau$, $C_{m, M}(\tau) < \infty$

2. $P(g_k(U) > \tau) \xrightarrow{k \to \infty} 1$. More precisely, if $\varepsilon \in [0, 1]$ then for all $k \geq k_{m, M}(\varepsilon, \tau)$, for all $p, q \in [m, M]$ we have

$$1 - \varepsilon < P(g_k(U) \geq \tau) \leq 1$$

3. $P(g_k(V) < \tau) \xrightarrow{k \to \infty} 1$. More precisely, if $\varepsilon \in [0, 1]$, then for all $k \geq k_{1-m, 1-M}(\varepsilon, \frac{1}{\tau})$, for all $p, q \in [m, M]$ we have:

$$1 - \varepsilon < P(g_k(V) < \tau) \leq 1.$$

: To prove the lemma we will use the Bernstein’s exponential inequality (for its proof, see for instance ?). Let $Z_1, ..., Z_n$ be iid real random variables with expectation $\mu$ and variance $\sigma^2$. Assume that there exists a finite constant $M$ such that $|Z_1 - \mu| \leq r$ a.s. Then, for any $\gamma > 0$:

$$P\left(\left|\frac{1}{n} \sum_{i=1}^{n} Z_i - \mu\right| \geq \gamma\right) \leq \exp\left(\frac{-n\gamma^2}{2(\sigma^2 + \frac{r}{4})}\right).$$

i) It is clear that $A$ and $C$ are continuous as functions of $p, q \in [m, M]$; therefore we only have to check that $A$ is strictly positive. This is equivalent to

$$\frac{\log\left(\frac{p}{1-q}\right)}{\log\left(\frac{(1-p)(1-q)}{pq}\right)} + (1 - p) > 0.$$

This happens iff

$$\frac{\log\left(\frac{p}{1-q}\right)}{\log\left(\frac{(1-p)(1-q)}{pq}\right)} > -(1 - p)$$

which is equivalent to

$$\log\left(\frac{p}{1-q}\right) + (1 - p) \log\left(\frac{(1-p)(1-q)}{pq}\right) > 0,$$

which in turn corresponds to

$$p \log\left(\frac{1-q}{p}\right) + (1 - p) \log\left(\frac{q}{1-p}\right) < 0.$$

Consider an auxiliary random variable $Z$ such that

$$P(Z = \frac{q}{1-p}) = 1 - p, \quad P(Z = \frac{1-q}{p}) = p,$$
then it is clear that $E(Z) = 1$ and, since $f(x) = \log x$ is a strictly concave function and $Z$ is not constant ($\frac{q}{1-p} < \frac{1}{2} < \frac{1-q}{p}$), then by Jensen inequality

$$p \log \left(\frac{1-q}{p}\right) + (1-p) \log \left(\frac{q}{1-p}\right) = E(\log(Z)) < \log (E(Z)) = 0$$

and therefore

$$A(p,q) > 0. \quad (14)$$

ii) Since

$$g_k(z) = \frac{P(U = z)}{P(V = z)} = \frac{(1-p)^zp^{k-z}}{q^z(1-q)^{k-z}} = \left(\frac{(1-p)(1-q)}{pq}\right)^z \left(\frac{p}{1-q}\right)^k,$$

then

$$g_k(z) \geq \tau \text{ iff } z \log \left(\frac{(1-p)(1-q)}{pq}\right) + k \log \left(\frac{p}{1-q}\right) \geq \log \tau.$$

Since $1-p > 1/2 > q$, $\log \left(\frac{(1-p)(1-q)}{pq}\right) > 0$. Thus

$$g_k(z) \geq \tau \text{ iff } z \geq \frac{\log \tau - k \log \left(\frac{p}{1-q}\right)}{\log \left(\frac{(1-p)(1-q)}{pq}\right)}.$$

Using this,

$$P(g_k(U) \geq \tau) = P \left( U \geq \frac{\log \tau - k \log \left(\frac{p}{1-q}\right)}{\log \left(\frac{(1-p)(1-q)}{pq}\right)} \right)$$

$$= P \left( \frac{U}{k} - (1-p) \geq \frac{C(p,q,\tau)}{k} - A(p,q) \right).$$

To apply Bernstein inequality, let us note that $U \sim Bin(k, 1-p)$ so that each of its components $T_j$ is such that $|T_j - (1-p)| \leq 1$, and $V(T_j) = p(1-p) \leq \frac{1}{4}$. If $k > \frac{2C_{m,M}(\tau)}{a_{m,M}}$ then

$$\frac{C(p,q,\tau)}{k} - A(p,q) < -\frac{a_{m,M}}{2}. \quad (15)$$

If in addition

$$k > \log \left(\frac{1}{\varepsilon}\right) \left(\frac{6 + 4a_{m,M}}{3a_{m,M}^2}\right)$$

by Bernstein inequality,

$$1 - P \left( \frac{U}{k} - (1-p) \leq -\frac{a_{m,M}}{2} \right) = P \left( \frac{U}{k} - (1-p) \geq -\frac{a_{m,M}}{2} \right) \leq \exp \left( -k \frac{(\frac{a_{m,M}}{2})^2}{2(\frac{1}{4} + \frac{1}{3} \frac{a_{m,M}}{2})} \right)$$
12  

and as $k \geq k_{m,M}(\varepsilon, \tau)$

$$P \left( \frac{U}{k} - (1 - p) \geq \frac{a_{m,M}}{2} \right) \leq \exp \left( -\log \left( \frac{1}{\varepsilon} \right) \left( \frac{6 + 4a_{m,M}}{3a_{m,M}^2} \right) \frac{3a_{m,M}^2}{6 + 4a_{m,M}} \right) = \varepsilon$$

Thus $P(g_k(U) \geq \tau) \geq 1 - \varepsilon$.

iii) For the last part of the Lemma, write down:

$$P(\hat{g}_k(V) < \tau) = P \left( \frac{1}{\hat{g}_k(V)} > \frac{1}{\tau} \right).$$

But

$$\frac{1}{g_k(z)} = \frac{P(V = z)}{P(U = z)} = \tilde{g}_k(z),$$

where $q$ plays the role of $1 - p$. Since $V \sim Bin(k, q)$, the results of i) apply if $[m, M]$ is replaced by $[1 - M, 1 - m]$. Then for

$$k \geq k_{1-m,1-M}(\varepsilon, \frac{1}{\tau})$$

one has $P(g_k(V) < \tau) > 1 - \varepsilon$. \hfill \Box$

We will now apply this Lemma to our results. Since $S_X$ is finite and (??) holds, from now on let

$$m = \min_{x \in S_X} \{\alpha(x), \beta(x)\} > 0$$

$$M = \max_{x \in S_X} \{\alpha(x), \beta(x)\} < 1/2$$

and

$$K(\varepsilon, \eta) = \max_{x \in S_X} \left\{ k_{m,M} \left( \varepsilon, \frac{p(0|x)(1 - \eta)}{p(1|x)\eta} \right), k_{1-M,1-m} \left( \varepsilon, \frac{p(0|x)\eta}{p(1|x)(1 - \eta)} \right) \right\} (16)$$

**Theorem 4.2:** Let $0 < \eta < 1$. Then:

1. $P(\hat{Y}_1^{(k)} = Y_1) \rightarrow 1$, 

2. $R_n^{(k)} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{\hat{Y}_i^{(k)} \neq Y_i\}} \xrightarrow{k \to \infty} 0$, and if $k \geq K(\varepsilon, \eta)$, then $\|R_n^{(k)}\|_1 < \varepsilon$

   To prove the first point we will show that, if $\varepsilon > 0$,
   then for all $k \geq K(\varepsilon, \eta)$, $P(\hat{Y}_1^{(k)} = Y_1) > 1 - \varepsilon$. Let us compute
\begin{align*}
P(\hat{Y}_1^{(k)} = Y_1) &= P(\hat{Y}_1^{(k)} = 1, Y_1 = 1) + P(\hat{Y}_1^{(k)} = 0, Y_1 = 0) \\
&= \sum_{x \in S_X} [P(\hat{Y}_1^{(k)} = 1, Y_1 = 1, X_1 = x) + P(\hat{Y}_1^{(k)} = 0, Y_1 = 0, X_1 = x)] \\
&= \sum_{z=0}^{k} \sum_{x \in S_X} P(\hat{Y}_1^{(k)} = 1, Y_1 = 1, X_1 = x, Z_1^{(k)} = z) \\
&\quad + \sum_{z=0}^{k} \sum_{x \in S_X} P(\hat{Y}_1^{(k)} = 0, Y_1 = 0, X_1 = x, Z_1^{(k)} = z)
\end{align*}

Replacing the first term in the sum by:

\[P(\hat{Y}_1^{(k)} = 1|Y_1 = 1, X_1 = x, Z_1^{(k)} = z)P(Z_1^{(k)} = z|Y_1 = 1, X_1 = x)p(1|x)p(x)\]

and the second by:

\[P(\hat{Y}_1^{(k)} = 0|Y_1 = 0, X_1 = x, Z_1^{(k)} = z)P(Z_1^{(k)} = z|Y_1 = 0, X_1 = x)p(0|x)p(x)\]

We have that:

\[P(\hat{Y}_1^{(k)} = Y_1) = \sum_{z=0}^{k} \sum_{x \in S_X} P(\hat{Y}_1^{(k)} = 1|Y_1 = 1, X_1 = x, Z_1^{(k)} = z)P(Z_1^{(k)} = z|Y_1 = 1, X_1 = x)p(1|x)p(x)\]
\[\quad + \sum_{z=0}^{k} \sum_{x \in S_X} P(\hat{Y}_1^{(k)} = 0|Y_1 = 0, X_1 = x, Z_1^{(k)} = z)P(Z_1^{(k)} = z|Y_1 = 0, X_1 = x)p(0|x)p(x)\]

We will apply the preceding Lemma using \(U = (Z_1^{(k)}|Y_1 = 1, X_1 = x) \sim \text{Bin}(k, 1 - \beta(x))\) and \(V = (Z_1^{(k)}|Y_1 = 0, X_1 = x) \sim \text{Bin}(k, \alpha(x))\). Recall that by definition of our MAP classifier,

\[P(\hat{Y}_1^{(k)} = 0|Y_1 = 0, X_1 = x, Z_1^{(k)} = z) = P \left[ \frac{P(U = z)p(1|x)}{P(V = z)p(0|x)} < \frac{\eta}{1 - \eta} \right] \]

So

\[\sum_{z=0}^{k} P(\hat{Y}_1^{(k)} = 0|Y_1 = 0, X_1 = x, Z_1^{(k)} = z)P(Z_1^{(k)} = z|Y_1 = 0, X_1 = x)\]

\[= \sum_{z=0}^{k} P \left[ \frac{P(U = z)p(1|x)}{P(V = z)p(0|x)} < \frac{\eta}{p(0|x)\eta} \right] P(V = z) \tag{17}\]

By Lemma 1 ii), given \(\epsilon \in [0, 1]\), if

\[\forall x \in S_X, k \geq k_{m,M}, \left( \epsilon, \frac{p(1|x)(1 - \eta)}{p(0|x)\eta} \right) \]
then
\[
\sum_{z=0}^{k} P(\hat{Y}_{1}^{(k)} = 0 | Y_1 = 0, X_1 = x, Z_1^{(k)} = z) P(Z_1^{(k)} = z | Y_1 = 0, X_1 = x) > (1 - \varepsilon)
\]
and
\[
\sum_{x \in S_X} \sum_{z=0}^{k} P(\hat{Y}_{1}^{(k)} = 0 | Y_1 = 0, X_1 = x, Z_1^{(k)} = z) P(Z_1^{(k)} = z | Y_1 = 0, X_1 = x) p(0|x)p(x)
\]
\[
> (1 - \varepsilon) \sum_{x \in S_X} p(0|x)p(x)
\]

In a similar way, if
\[
\forall x \in S_X, k \geq k_1 - M, 1 - m \left( \varepsilon, \frac{p(0|x)\eta}{p(1|x)(1 - \eta)} \right)
\]
then,
\[
\sum_{z=0}^{k} \sum_{x \in S_X} P(\hat{Y}_{1}^{(k)} = 1, Y_1 = 1, X_1 = x, Z_1^{(k)} = z) P(Z_1^{(k)} = z | Y_1 = 0, X_1 = x) p(1|x)p(x)
\]
\[
> (1 - \varepsilon) \sum_{x \in S_X} p(1|x)p(x)
\]

That is for all \( k \geq K(\varepsilon, \eta) \) we have:
\[
P(\hat{Y}_{1}^{(k)} = Y_1) > \sum_{x \in S_X} (1 - \varepsilon)p(0|x)p(x) + \sum_{x \in S_X} (1 - \varepsilon)p(1|x)p(x) = 1 - \varepsilon,
\]
thus, i) is proved.

ii) Observe that
\[
R_n^{(k)} \geq 0,
\]
\[
E(R_n^{(k)}) = P(\hat{Y}_{1}^{(k)} \neq Y_1) \underset{k \to \infty}{\longrightarrow} 0,
\]
and
\[
\|R_n^{(k)}\|_1 < \varepsilon
\]
if \( k \geq K(\varepsilon, \eta) \).

Next result shows the asymptotic behavior of our algorithm when \( n \) goes to infinity.

**Corollary 4.1:**

With the same notation as above, if \( k \geq K(\varepsilon, \eta) \) then
\[
P(\limsup_n R_n^{(k)} \geq \varepsilon) = 0.
\]
Since \( \hat{Y}_i^{(k)} \) is a function of \( X_i \) and \( Z_i^{(k)} \), which is itself a given function of \( T^{(1)}(X_i, Y_i), \ldots, T^{(k)}(X_i, Y_i) \), \( 1_{\{\hat{Y}_i^{(k)} \neq Y_i\}} \) are iid and hence, by the Strong Law of Large Numbers,

\[
P^{(k)}_n \xrightarrow{a.s.} n \to \infty P(\hat{Y}_1^{(k)} \neq Y_1)
\]

Applying Theorem 2, we get the result.

\[\square\]

### 5 Simulations

In order to validate our classification procedure we tried several simulations. The explanatory variables are binary \( d \)-dimensional, \( X \in S_X = \{0, 1\}^d \). The output \( Y \in \{0, 1\} \) is such that:

\[P[Y = 1] = 0.95 \text{ if } \sum_{j=1}^d x_j > 6\]

and

\[P[Y = 0] = 0.95 \text{ if } \sum_{j=1}^d x_j \leq 6\]

For each \( x = (x_1, \ldots, x_d) \in S_X \) we set:

\[
\alpha(x) = \frac{1}{4d} \sum_{j=1}^d 2^{x_j}, \quad \beta(x) = \frac{1}{4d} \sum_{j=1}^d x_j
\]

\( \alpha(x) \) and \( \beta(x) \) are < \( \frac{1}{2} \). For \( j = 1, 2, \ldots, k \), we sample \( T^{(j)}(x, y) \) using a Bernoulli distribution of parameter \( \alpha(x) \) when \( y = 0 \), and \( 1 - \beta(x) \) when \( y = 1 \). The statistics \( \pi_n \) and \( N_i^{(j)} \) are computed as in (??). We apply then our algorithm described in table ?? twice, once using the true values of the parameters \( \alpha(x), \beta(x), \) and \( p(1|x) \), and once using their estimates.

Three factors are involved in our simulations: the sample size \( n \), the number of iterations \( k \), and the number of equations \( neq \) used in the non linear system. We set \( d = 11, \quad k = 2000, \quad \eta = 0.5, \quad neq = 3, 5, 10, 20 \) and \( n = 20000, 50000, 100000, 200000 \). It should be noticed that as \( X \) is discrete and all its components are binary, \( \alpha(x) \) and \( \beta(x) \) are discrete, and the number of different observations is \( 2^p \). Extreme values of these functions are observed when all \( X \) components are equal to zero or one. As this is rare, in these cases estimates of \( \alpha \) and \( \beta \) are poor as shown in figures ?? and ??.
Figure 1  Estimated versus observed values of $\alpha$. 
Bayesian Learning of unobservable output

Figure 2  Estimated versus observed values of $\beta$.

Figure ?? shows how the mean classification error of our algorithm evolves within the number of iterations. The dashed lines correspond to the error computed when we use in our classifier the estimated values of the unknown parameters, and the solid line correspond to the use of the true values of these parameters. The vertical columns correspond to variation in the sample size $n$, and the horizontal ones for different values of the number of equations $neq$. For clarity we have omitted the graphics for the highest values of $n$ and $neq$.

Table ?? gives the minimum misclassification error rates achieved in the simulations. We get a smaller error rate when increasing the sample size and the number of equations used to estimate the parameters.

<table>
<thead>
<tr>
<th>$n/neq$</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>20000</td>
<td>0.26</td>
<td>0.25</td>
<td>0.23</td>
<td>0.2</td>
</tr>
<tr>
<td>50000</td>
<td>0.23</td>
<td>0.21</td>
<td>0.18</td>
<td>0.16</td>
</tr>
<tr>
<td>100000</td>
<td>0.22</td>
<td>0.18</td>
<td>0.14</td>
<td>0.13</td>
</tr>
<tr>
<td>200000</td>
<td>0.2</td>
<td>0.15</td>
<td>0.11</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Table 2  Minimum mean classification error of our algorithm using $k = 2000$ iterations, and varying sample size $n$ and the number of equations $neq$ used to solve the non linear system and estimate the parameters.

6 Discussion

In this paper we have proposed a bayesian algorithm to predict a non observable output using a sequence of tests which may be seen as weak classifiers whose
estimation does not utilize $Y$ as it is not available. We have demonstrated the efficiency of our classifier both theoretically and using simulations. As no other algorithms are available for our problem we compared our approach to more classical learning techniques as in the simulations the true values of the output variable are known. Although our algorithm may be used as is in real applications we could not yet get a real data set for a real application. Finally, although we focused on the binary case, it should be mentioned that our algorithm may be directly generalized for the multiclass case.
References


