Boolean regression*

E. Boros

DIMACS and RUTCOR, Rutgers University,
P.O. Box 5062, New Brunswick, NJ 08903, USA

P.L. Hammer

RUTCOR, Rutgers University,
P.O. Box 5062, New Brunswick, NJ 08903, USA

J.N. Hooker

GSIA, Carnegie-Mellon University, Pittsburgh, PA 15213, USA

We take a regression-based approach to the problem of induction, which is the problem of inferring general rules from specific instances. Whereas traditional regression analysis fits a numerical formula to data, we fit a logical formula to boolean data. We can, for instance, construct an expert system for fitting rules to an expert's observed behavior. A regression-based approach has the advantage of providing tests of statistical significance as well as other tools of regression analysis. Our approach can be extended to nonboolean discrete data, and we argue that it is better suited to rule construction than logit and other types of categorical data analysis. We find maximum likelihood and bayesian estimates of a best-fitting boolean function or formula and show that bayesian estimates are more appropriate. We also derive confidence and significance levels. We show that finding the best-fitting logical formula is a pseudo-boolean optimization problem, and finding the best-fitting monotone function is a network flow problem.

1. Introduction

A number of practical situations require that we identify a boolean function that captures a relationship implicit in a set of observations. We observe the value of an unknown function for certain values of its arguments, and we wish to infer, as best we can, what the function is. A *boolean* function is appropriate when the function and its arguments take only two values. We begin with some examples.

*The first and second authors gratefully acknowledge the partial support of NSF (Grant DMS 89-06870) and AFOSR (Grants 89-0512 and 90-0008), and the third author that of AFOSR (Grant 91-0287) and ONR (Grant N00014-92-J-1028).

1.1. SOME EXAMPLES

One class of examples arises from efforts to build a rule base for an expert system. Suppose we have records of a bank officer's past decisions when he was presented with loan applications. We would like to use these as a basis for formulating a rule that indicates when loans should be granted. Each application is characterized by the presence or absence of a fixed set of attributes (e.g. whether the applicant is employed, whether he has a good credit record, etc.) and by the loan officer's decision. We allow for some noise in the data, since the officer may not always decide the same way when presented with applications having the same profile. The data are also incomplete in the sense that we do not have an application representing every possible set of characteristics. The rule we want to derive takes the form of a boolean function whose arguments indicate, by 0 or 1, whether each attribute is present, and whose value indicates whether the loan should be granted.

A similar approach can be taken to deciding whether to audit an income tax form, whether to admit a patient into the hospital, whether to replace an electronic component, whether to grant someone government benefits, and so on.

A related class of applications involve the *prediction* of a boolean outcome on the basis of boolean data. We may, for instance, wish to predict whether a substance is carcinogenic on the basis of some tests we perform on it [3]. Each test has a boolean outcome: "positive" or "negative". The data set consists of test results for a number of substances that were also investigated clinically for carcinogenicity (again a boolean outcome). Noisy data are possible, since the clinical trials could be misleading. The object is to discover the boolean function that predicts, on the basis of the test results, whether a substance is carcinogenic.

In a third class of applications, we want to *explain* an effect. A physician, for instance, may want to determine which combinations of foods cause his patient's suspected food allergy [5]. He asks the patient to record, each day, whether he eats certain foods, and whether an allergic reaction develops. The object is to find the boolean function that relates the occurrence of a reaction to which foods were eaten.

1.2. ADVANTAGES OF BOOLEAN REGRESSION

We propose a regression-based approach to discerning a boolean relationship. That is, we propose to fit a boolean function to the data in much the same way that classical regression analysis fits a numerical function to data. Our basic motivation is that a regression approach permits us to analyze statistical significance, so that genuine relationships (those likely to appear in any sample of observations) can be distinguished from spurious ones (those that are artifacts of the sample). The computational problems posed by the boolean regression problem are quite different from those of classical regression, however, and we will show how to attack them with the methods of combinatorial optimization.

The problem we address can be viewed as a classification problem because a boolean function $f(x)$ in effect divides values of x into two classes: those for which $f(x) = 1$ and those for which $f(x) = 0$. The literature of pattern recognition and machine learning offers a wide variety of classification techniques based on clustering, classification trees, discriminant functions, and so on [4, 10, 12, 13]. However, the regression approach we propose has the distinct advantage that it, like classical regression analysis, can test the statistical significance of the relationship discovered and the degree of confidence one can have that it is correct. It can also provide a logical formula for predicting an outcome, together with an analysis of the statistical significance of terms in the formula. Finally, it makes available stepwise regression and other regression-based techniques for discovering significant relationships.

It is true that logit analysis, a well-known type of regression analysis designed for boolean data, also provides confidence and significance levels. (Logit analysis is a special case of categorical data analysis in which the data are boolean and there is only one dependent variable [1,2,7].) However, a logit model is in fact quite different from our boolean regression model. A logit model predicts the *probability* that $f(x) = 1$, whereas our model predicts *whether* $f(x) = 1$. It also assumes that the probability that $f(x) = 1$ is given by a log-linear function of x, whereas we make no such assumption.

A logit model is appropriate when we are genuinely interested in the probability that $f(x) = 1$, along with the level of confidence we can have in the probability estimate. Logit models are commonly used, for instance, to predict consumer demands. Here, $f(x) = 1$ means that a consumer having profile x will buy product A. We can predict how many consumers having profile x will buy product A by multiplying the market population by the predicted probability that $f(x) = 1$.

However, in many applications we want to know whether $f(x)$ is 1 or 0, along with the level of confidence we can have in the prediction. We may wish to know whether a loan should or should not be granted, or whether a substance is or is not carcinogenic. Furthermore, if we determine that a loan should be granted, we may wish to know what confidence we can have in *this determination*, rather than a confidence interval for some estimate of the probability that it should be granted. We may also want a logical formula that instructs us when to grant a loan, rather than a log-linear or any other numerical formula.

The approach we describe can in principle be extended to nonboolean discrete functions. The loan applicant's credit rating, for instance, may be "poor", "acceptable", or "excellent", and the bank officer's decision may be "reject", "accept", or "get more information". In such cases, one can encode many-valued attributes using two or more boolean variables, and one can develop a different boolean function for each possible outcome. The resulting predictions, however, are not necessarily the same that result from using a single nonboolean function. In [3], we take a first step toward solving the nonboolean case: we show how to find an error-minimizing fit with a *monotone* nonboolean function by solving an easy network flow problem, provided the possible values of the function have an interval order. However, nonboolean functions need further investigation.

1.3. OUTLINE OF THE PAPER

In section 2, we begin with the basic concepts of boolean regression. We introduce the regression problem with an example to which we refer throughout the paper. We note that practical application generally requires that the regression function have a certain form or certain properties; i.e. that it belongs to a specified class of boolean functions.

In section 3, we show how to find the best fitting function f that belongs to any specified class F of boolean functions. We first derive maximum likelihood estimates (MLEs) of f and show that such an estimate has the curious property that it either minimizes or maximizes the number of errors. We note that an error-maximizing estimate may be suitable for some applications but not others. We therefore develop a more general bayesian approach that allows one to obtain an MLE if desired but can deal with situations in which error-maximizing MLE is inappropriate. The bayesian approach also provides a natural way to derive confidence intervals and significance levels. In particular, we show how to derive bayesian estimates, as well as confidence and significance levels, under two practical models of the prior distribution of error probabilities. We do so in such a way that only an error-minimizing f and an error-maximizing f need be computed, since these computations are relatively easy to perform. Again, these results are valid for any specific class F within which the fitted function is required to lie.

The next three sections show how to do the computations more efficiently for three particular types of function classes F. Section 4 treats the case in which F contains *all* boolean functions; that is, the fitted function f may be any boolean function.

In section 5, F is assumed to contain all *monotone* boolean functions. The loan officer's function, for instance, is likely to be monotone, since if the presence of certain positive attributes indicates that the loan should be granted, then the presence of these and still other positive attributes certainly indicates that the loan should be granted. We show that the error-minimizing monotone fit can be quickly obtained with a minimum cut computation in a network flow problem. This result is generalized in [3].

In section 6, F is assumed to contain all boolean functions *defined by logical formulas having a specified form.* This is perhaps the case most reminiscent of classical numerical regression, since it supposes that the true relationship can be expressed as a formula of propositional logic having a certain form, just as in classical regression one might suppose that the relationship is linear. However, rather than estimate the value of numerical coefficients, we use regression techniques to find which terms of the formula should be included and which omitted. We also indicate how to check whether the inclusion of a particular term or set of terms results in a statistically significant improvement in the fit. We show that the problem of computing the best fit can be formulated as a pseudo-boolean optimization problem, which is well studied in the operations research literature. We also provide means of approximating confidence and significance levels.

2. Basic concepts

2.1. THE BOOLEAN REGRESSION MODEL

Let y be a boolean variable that is dependent on a vector $x = (x_1,\dots,x_n)$ of boolean variables. For instance, if the loan officer mentioned earlier wishes to decide whether to make a loan, the variables $x_1,\dots,x_n$ would describe the circumstances and y the officer's decision. Perhaps $x_1 = 1$ indicates that the applicant is now paying off a mortgage loan, $x_2 = 1$ indicates that he or she has a good credit record, and so on. If $y = 1$, the officer decides to lend the money, and $y = 0$ otherwise.

We would like to capture the loan officer's past decisions in a boolean function f that indicates the correct decision $f(x)$ for any vector x. Suppose we have observed a series of pairs (x, y), such as those displayed in table 1. Note that each observed set of circumstances arose several times, and the loan officer's decisions were not consistent. For instance, in the set of circumstances denoted by $x = (0, 1, 0, 1, 1)$, the officer granted the loan in fifteen cases and refused it in the other seven. Also, only six of the $2^5 = 32$ possible sets of circumstances were observed.

Table 1

A sample data set.

Circumstances					No. of observations with	
x_1	x_2	x_3	x_4	x_5	$y=0$	$y=1$
0	1	0	1	1	7	15
1	1	0	0	1	9	3
1	0	0	1	1	8	2
1	0	0	0	1	3	7
1	1	0	1	0	5	17
0	0	1	1	1	9	2

To derive a regression model, we will assume that *some* boolean function of variables $x_1,\dots,x_5$ is a true description of the loan officer's judgment. Since no boolean function exactly fits the observations, due to the officer's inconsistency, we must attribute that inconsistency to "random error" of some sort.

This raises the question as to what sort of random variable best models observational error. An ordinary numerical regression is written as

$$y = g(x) + \varepsilon, \tag{1}$$

where g is a numerical function of a vector x of independent variables that take numerical values. The error term ε is classically assumed to have a normal distribution with mean zero. When this assumption is unwarranted, various nonparametric methods are used.

In the boolean case, the situation is actually much simpler, since there are only two possible errors: the outcome is 1 when it should be 0, and vice versa. We therefore write the regression as follows:

$$y = f(x) \oplus \varepsilon. \tag{2}$$

Here, $\oplus$ is a binary sum, so that $a \oplus b = (a + b) \bmod 2$. We let the boolean error term ε be a simple Bernoulli variable: it takes the value 1 with probability p and the value 0 with probability $1 - p$. Our distributional assumptions are therefore quite minimal, and it is difficult to see how a "nonparametric" model could assume less.

It may sometimes be useful, however, to generalize the error model in one respect. We may want to use an *asymmetric* error model in which the probability of observing $y = 0$ when $f(x) = 1$ differs from the probability of observing $y = 1$ when $f(x) = 0$. In principle, the ensuing analysis can be carried out for an asymmetric error model, but since the results are substantially more complex, we treat it only briefly in section 3.4.

2.2. PARTIALLY AND FULLY DEFINED FUNCTIONS

Boolean regression predicts the value of a boolean function only for *observed* values of x. This is because it finds a best-fitting *partially defined* function f, in particular one that is defined only for the observed values of x. To predict the value of the function for unobserved x, one must *extend* f to a function f^* that is defined on all x. That is, one must find a fully-defined function f^* that agrees with f on observed x.

From here on, we will let f denote the partially defined function sought by boolean regression. If the extended function f^* is required to belong to a class F^* of boolean functions, we let F denote the class of partially defined boolean functions with extensions in F^*. For instance, if we require the extended function f^* to be monotone, then F is the class of monotone functions that are defined only on the observed values of x.

If no restrictions are placed on f^*, then the value of $f^*(x)$ is completely arbitrary for unobserved x, and boolean regression makes no predictions for these values. Thus, much of the practical value of regression is lost. Furthermore, the solution of the regression problem is trivial: for each observed x, we can let $f(x)$ be the value most often observed for y. There is no point in solving even this trivial problem unless y is observed several times for each observed x, a condition often not met in practice.

When f^* is required to belong to a particular class F^* of boolean functions, however, the predicted values of $f(x)$ for observed x may indirectly determine the value of the extended function $f^*(x)$ for many unobserved x. This is because all the extensions of f in F^* may all agree at many unobserved values of x. Furthermore, when f is so restricted, the boolean regression problem is no longer trivial, and as in classical regression, there is no need to collect multiple observations of y for a given x.

We will therefore analyze the regression problem under the general assumption that the fitted function f is required to belong to some prespecified class F of partially defined boolean functions. This is also more in tune with the spirit of classical regression analysis, in which the form of the regression function (e.g. linear) is chosen to suit the application. It may, for instance, be dictated by a theory of hypothesis about the relation the regression is intended to capture.

We will pay particular attention to the cases in which f is required to be monotone or to be defined by a given type of logical formula, because of their practical importance and their computational advantages. In the former case, we will indicate how to find the fitted f and extend it to a fully defined f^*. In the latter case, we suppose that the regression function is defined by a logical formula having a certain form, and we derive which formula of this form is the best fit.

One simple example of a functional form is

$$f(x) = \beta_0 \vee \beta_1 x_1 \vee \ldots \vee \beta_n x_n, \tag{3}$$

where "$\vee$" means "or". Here, the β_j's might be boolean coefficients that are to be estimated, with the understanding that x_j appears in the formula if $\beta_j = 1$ and is absent if $\beta_j = 0$. (Some useful functional forms will be discussed in section 6.)

Since a logical formula has a well-defined value for any given x, the best-fitting formula defines the extended function f^* as well as f. If more than one formula fit the data equally well, $f^*(x)$ is determined for unobserved x when these formulas agree.

3. The general regression problem

3.1. MAXIMUM LIKELIHOOD ESTIMATION

We now address the general question as to how to find a best-fitting boolean function f when f must belong to an arbitrary class F of partially defined boolean functions. In this section, we derive an MLE estimate and in the next section, develop a bayesian approach. In sections 3.5 and 3.6, we use the bayesian model to derive statistical tests for boolean regression.

The most natural estimate of f is the function in F that agrees with observation most often. This would be the function that minimizes the number of errors, which is the number of observations for which $f(x)$ and y differ. We will see, however, that the error-minimizing function is often but not always an MLE. We will prove the somewhat surprising fact that an MLE of f is *either error minimizing* or *error maximizing*.

A *maximum likelihood estimate* (*MLE*) of a function value (or parameter) is one that maximizes the probability of making the observations y that were in fact made. In the classical regression model (1), the MLE is obtained by the well-known least-squares calculation, provided the errors are independently and normally distributed.

We can illustrate maximum likelihood estimation using the example of table 1. Let $X = \{x^1,\dots,x^6\}$ be the set of observed values of x. We wish to find the values $f(x^1),\dots,f(x^6)$ that maximize the likelihood of making the observations in table 1. Since the likelihood of making these observations depends on the probability p of error, we must estimate p and the $f(x^i)$'s simultaneously. (In classical regression, one simultaneously estimates the functional parameters and the variance of the error distribution.)

If $(f(x^1),\dots,f(x^6)) = (0, 0, 0, 0, 0, 0)$, the likelihood of making the observations in table 1 is the likelihood of making seven correct and fifteen erroneous observations of $f(0, 1, 0, 1, 1)$, nine correct and three erroneous observations of $f(1, 1, 0, 0, 1)$, and so on, for a total of forty-one correct and forty-six erroneous observations. This likelihood is $(1-p)^{41}p^{46}$. Thus, we wish to compute

$$\max\{\max_p\{(1-p)^{41}p^{46},(1-p)^{34}p^{53},\dots,(1-p)^{46}p^{41}\}\}, \tag{4}$$

where the first term corresponds to $(f(x^1),\dots, f(x^6)) = (0, 0, 0, 0, 0, 0)$, the second to $(0, 0, 0, 0, 0, 1)$, and so on to $(1, 1, 1, 1, 1, 1)$, for a total of $2^5 = 32$ terms.

More generally, if Y represents the set of observations actually made, we want to compute

$$\max_{0\le p\le 1, f\in F} Pr(Y\,|\,f,p), \tag{5}$$

where $Pr(Y\,|\,f,p)$ is the likelihood function, given by

$$Pr(Y\,|\,f,p) = (1-p)^{N-e(f)}p^{e(f)}. \tag{6}$$

Here, N is the total number of observations and $e(f)$ is the number of erroneous observations for f, so that in table 1, $e(f) = 46$ when $(f(x^1),\dots,f(x^6)) = (0, 0, 0, 0, 0, 0)$.

We first observe that (6) is maximized with respect to p when $p = e(f)/N$. We can therefore replace every p in (6) with $e(f)/N$ and maximize the resulting expression, namely

$$\left[1-\frac{e(f)}{N}\right]^{N-e(f)}\left[\frac{e(f)}{N}\right]^{e(f)}, \tag{7}$$

over all f in F. Furthermore, due to the convexity of (7) in $e(f)$, it is clear that (7) is largest when $e(f)$ takes its largest or smallest possible value. These correspond, respectively, to an error-minimizing estimate $\hat{f}_1$ and an error-maximizing estimate $\hat{f}_2$ of f. Also, since (7) takes the same value when $e(f)$ is replaced by $N - e(f)$, setting $f = \hat{f}_1$ makes (7) larger than setting $f = \hat{f}_2$ when $e(\hat{f}_1) < N - e(\hat{f}_2)$. We therefore have the following.

THEOREM 1

Let $\hat{f}_1$ and $\hat{f}_2$, respectively, be error-minimizing and error-maximizing estimates of a boolean function f defined on X, subject to the condition that f belong to a given set F of boolean functions defined on X. Then either $\hat{f}_1$ or $\hat{f}_2$ (possibly both) is an MLE $\hat{f}$ of f, and $e(\hat{f})/N$ is the corresponding MLE of the error probability p. In particular, $\hat{f}_1$ is an MLE of f if and only if $e(\hat{f}_1) \leq N - e(\hat{f}_2)$, and $\hat{f}_2$ is an MLE if and only if $e(\hat{f}_1) \geq N - e(\hat{f}_2)$.

If no restrictions are placed on f, the error-minimizing function $\hat{f}_1$ is obviously the one that minimizes the number of errors for each observed x, and analogously for the error-maximizing function $\hat{f}_2$. Clearly, $e(\hat{f}_1) = N - e(\hat{f}_2)$, so that both are MLEs.

COROLLARY 1

If no restrictions are placed on f, then the error-minimizing estimate $\hat{f}_1$ given by

$$\hat{f}_1(x) = \begin{cases} 0 & \text{if } y = 0 \text{ was observed more often than } y = 1, \\ 1 & \text{otherwise,} \end{cases} \tag{8}$$

is a maximum likelihood estimate of f, and $\hat{p}_1 = e(\hat{f}_1)/N$ is the corresponding estimate of the error probability. The error-maximizing function $\hat{f}_2 = 1 - \hat{f}_1$ is also an MLE, with an error probability estimate of $1 - \hat{p}_1$.

If we suppose that the above example places no restrictions of f, the error-minimizing fit is given by $(\hat{f}_1(x^1),\ldots,\hat{f}_1(x^6)) = (1, 0, 0, 1, 1, 0)$. For this model, the number of errors is $e(\hat{f}_1) = 22$, and the estimated probability of error is $\hat{p}_1 = 22/87 = 0.253$. The error-maximizing estimate is the complement of $\hat{f}_1$, with a corresponding estimated probability 65/87 of error.

It is important to note that $\hat{p} = e(\hat{f})/N$ is a biased estimator of p. To see this, let N_x be the number of observations of $f(x)$, and let $b_p(N, k)$ be the binomial probability

$$b_p(N,k) = \binom{N}{k} p^k (1-p)^{N-k}. \tag{9}$$

Then if p is the true probability of error, the expected value of its estimate is

$$\begin{aligned} E(\hat{p}) &= \frac{1}{N} \sum_{x \in X} \left(\sum_{k=0}^{[N_x/2]} k b_p(N_x,k) + \sum_{k=[N_x/2]+1}^{N_x} (N_x - k) b_p(N_x,k) \right) \\ &= p - \frac{1}{N} \sum_{x \in X} \sum_{k=[N_x/2]+1}^{N_x} (2k - N_x) b_p(N_x,k). \end{aligned} \tag{10}$$

Here, $[N_x/2]$ is the largest integer less than or equal to $N_x/2$. In the example, if $p = 0.253$ were the true error probability, $E(\hat{p})$ would be 0.250, so the bias is minimal in this case. When there are fewer observations of each $f(x)$, however, bias can be significant.

3.2. WHEN ERROR MAXIMIZATION IS APPROPRIATE

In boolean regression, an MLE can maximize error. For the sample data set of table 1, for example, both the error-minimizing and error-maximizing functions are MLEs. In such a case, we are strongly tempted to discard the error-maximizing function as a spurious solution. Presumably, this is because we consider a larger error probability to be less likely, in some sense, than a smaller one.

It is not so clear what we should do, however, when the error-maximizing solution results in a strictly greater likelihood than the error-minimizing solution. (This can happen when f is restricted.) There are some situations in which the error-maximizing solution seems appropriate, whereas others seem to call for an error-minimizing solution.

Consider an example in which we have synthesized m very similar drugs and would like to put the purest one on the market. We checked whether each drug contains a certain impurity, using a test of questionable reliability, and the test failed to detect an impurity in any of the drugs. The chemistry of the production process, however, indicates that either the second half of the varieties are impure, or else all the varieties but the first are impure, and we would like to know which is the case. (Assume for convenience that m is even.)

Let the boolean vectors $x^1,\ldots,x^m$ describe the drugs, and let f be the unknown function of x that takes the value 1 when x describes an impure drug. We made one observation, namely $y = 0$, for each x^i. We know that the set F of possible functions contains two functions: the function f_1 that makes half the drugs impure, given by $f_1(x^i) = 0$ for $i = 1,\ldots,m/2$ and $f_1(x^i) = 1$ for $i = (m/2) + 1,\ldots,m$, and the function f_2 that makes all but one drug impure, given by $f_2(x^1) = 0$ and $f_2(x^i) = 1$ for $i = 2,\ldots,m$. The estimated error probabilities for the two functions, respectively, are $\hat{p}_1 = 1/2$ and $\hat{p}_2 = (m-1)/m$, and the corresponding likelihoods are $Pr(Y\,|\,\hat{f}_1, \hat{p}_1) = (1/2)^m$ and $Pr(Y\,|\,\hat{f}_2, \hat{p}_2) = (1/m)[(m-1)/m]^{m-1}$. For instance, if there are ten drugs ($m = 10$), the estimated error probabilities are 0.5 and 0.9, whereas the likelihoods are 0.00098 and 0.03874. So, concluding that half the drugs are impure results in about half as many errors as concluding that all but one are impure, but it makes the observation set forty times less likely. It may therefore be more reasonable to say that the test was almost always wrong than to say it was wrong half the time.

In another situation, however, the same data could lead to an opposite conclusion. Suppose that instead of drugs we have m applicants for a loan, and the loan officer rejected every one. Perhaps there are two schools of lending, one of which would recommend accepting only half the applications (function f_1), and one of which would recommend accepting all but one application (function f_2). It would be very odd to conclude that the loan officer belongs to the more lenient school but almost always makes mistakes, rather than concluding that he is strict and makes mistakes only half the time, even though the former hypothesis makes the observation set much more likely.

The difference between the two scenarios might be explained as a difference in the assumed prior probability distribution for the error probability p. In the drug test, the distribution may be close to uniform over the unit interval, since we have limited information on the reliability of the test. A loan officer, however, is much more likely to make a few mistakes than many mistakes. This suggests that if the MLE maximizes error, it may be appropriate to use a bayesian approach that begins with a reasonable prior distribution of error probabilities [15]. A prior distribution that favors smaller values of P may result in an error-minimizing estimate.

3.3. BAYESIAN ESTIMATION

In the bayesian approach, we suppose that the error probability p has a prior probability distribution with a known density function π, and that each $f \in F$ has some known prior probability $Pr(f)$. We will assume that the prior distributions of f and p are independent. Let Y represent the observation set that was actually gathered. We wish to know the posterior probability (density) of any given pair (f, p), given that observations Y were made. If we let $h(f, p \mid Y)$ represent this density function, then Bayes' rule says

$$h(f, p \mid Y) = \frac{Pr(Y \mid f, p)Pr(f)\pi(p)}{Pr(Y)}, \tag{11}$$

where $Pr(Y \mid f, p)$ is given by (6). The denominator of the fraction is the marginal probability of Y, given by

$$Pr(Y) = \sum_{f' \in F} \int_0^1 Pr(Y \mid f', p')Pr(f')\pi(p')dp'. \tag{12}$$

Our bayesian estimates $\tilde{f}, \tilde{p}$ will be those that *maximize posterior probability*, i.e. a pair (f, p) that solves

$$\max_{0 \le p \le 1, f \in F} Pr(Y \mid f, p)Pr(f)\pi(p). \tag{13}$$

It is clear from (13) that if the prior probabilities of f and p are uniformly distributed (so that $Pr(f)$ and $\pi(p)$ are constant), this bayesian estimate is identical to the MLE.

Thus, if no MLE is error minimizing (or, as we will see, one wishes to obtain confidence and significance levels), then it is necessary to think about the prior distribution of p. We will consider two probability models that may be useful in practice. Both assume that all $f \in F$ have equal prior probability, so that $Pr(f) = 1/|F|$. We will describe how to use the models in such a way that one need only compute error-minimizing and error-maximizing functions.

It will be convenient to define the following estimates $\hat{f}_i$, where $\hat{f}_1$ and $\hat{f}_2$ are the error-minimizing and error-maximizing estimates defined earlier. $\hat{p}_i$ is the corresponding estimate of error $e(\hat{f}_i)/N$.

$$\begin{aligned} \hat{f}_1 &= \text{a function } f \text{ that minimizes } e(f); \\ \hat{f}_2 &= \text{a function } f \text{ that maximizes } e(f); \\ \hat{f}_3 &= \text{a function } f \text{ that maximizes } e(f) \text{ subject to } \hat{p}_3 \le \rho. \end{aligned}$$

MODEL 1

The prior distribution of p is uniform on the interval $[0, \rho]$.

This would be appropriate when essentially nothing is known about the distribution of p except that it would be absurd to suppose $p > \rho$. (If nothing at all is known, $\rho = 1$, and we use the MLE.) For example, it may be unreasonable to explain an expert's behavior with a model on which he deviates from his own guidelines most of the time; in this case, $\rho = 1/2$. The distribution function becomes $\pi(p) = 1/\rho$ for $0 \le p \le \rho$, and $\pi(p) = 0$ elsewhere. The bayesian estimate $(\tilde{f}, \tilde{p})$ is obtained simply by solving

$$\max_{0 \le p \le \rho,\, f \in F} Pr(Y \mid f, p). \tag{14}$$

Theorem 1 and the convexity of (7) immediately imply the following.

COROLLARY 2

Let $\hat{f}$ be an MLE and $\hat{p}$ the corresponding error probability. Assume $\hat{p}_1 \le \rho$. Then the bayesian estimate $\tilde{f}$ under Model 1 is $\hat{f}_1$ if $\hat{f}$ is error minimizing ($\hat{f} = \hat{f}_1$). If $\hat{f}$ is not error maximizing, then $\tilde{f}$ is either $\hat{f}_1$ or $\hat{f}_3$. In either case, the estimate or error probability is $\tilde{p} = e(\tilde{f})/N$.

Because $\hat{f}_3$ may be hard to compute, it may not be practical to use Model 1 when the MLE is error maximizing.

MODEL 2

The probability density of p belongs to a family of functions given by $\pi(p) = (1 + \kappa)(1 - p)^{\kappa}$.

Thus, if $\kappa = 0$ the distribution is uniform, and if $\kappa = 1$ the probability of p decreases linearly as p increases. Since $Pr(Y|f, p)$ in (13) is given by (6), the bayesian estimate is obtained by solving

$$\max_{0 \le p \le 1, f \in F} (1+\kappa)(1-p)^{N+\kappa-e(f)} p^{e(f)}. \tag{15}$$

This has the same solution as the maximum likelihood problem (5), except that N is replaced by $N+\kappa$. Thus, theorem 1 implies the following.

COROLLARY 3

Let $\hat{f}$ be the MLE and $\hat{p}$ the corresponding error probability. Then the bayesian estimate $\tilde{f}$ under Model 2 is given by

$$\tilde{f} = \begin{cases} \hat{f}_1 & \text{if } e(\hat{f}_1) \le N+\kappa-e(\hat{f}_2), \\ \hat{f}_2 & \text{if } e(\hat{f}_2) \le N+\kappa-e(\hat{f}_1). \end{cases}$$

The corresponding estimate of error probability is $\tilde{p} = e(\tilde{f})/(N+\kappa)$.

One can always choose a κ large enough (i.e. a κ that makes large p's unlikely enough) so that the bayesian fit f is error minimizing. However, as we noted earlier, there are situations in which an error-maximizing fit may be appropriate. In the example involving the impure drugs and the loan officer, Model 2 picks the error-minimizing solution when

$$e(\tilde{f}_1) = m/2 \le m + \kappa - (m-1) = N + \kappa - e(\tilde{f}_2),$$

or when $\kappa \ge (m/2) - 1 = 4$ (using $m = 10$). Thus, it is more reasonable to conclude that the loan officer is wrong only half the time (rather than almost all of the time) when $\kappa \ge 4$. This corresponds to a prior distribution in which there is at least a 97% chance that the officer's probability of error is less than 1/2; i.e. we are 97% confident that the officer is right at least half the time. Conversely, we should conclude that all but one of the drugs is impure when $\kappa \le 4$. This occurs when there is at least a 3% chance that the test's reliability is less than 1/2; i.e. we admit a 3% possibility that the test is wrong at least half the time. These results do not seem unreasonable.

3.4. AN ASYMMETRIC ERROR MODEL

Let us suppose for the moment that we have asymmetric error probabilities. That is, the probability p_0 of observing $y = 0$ when $f(x) = 1$ is possibly different from the probability p_1 of observing $y = 1$ when $f(x) = 0$. In the preceding discussion, $p_0 = p_1 = p$. To obtain the likelihood function $Pr(Y \mid f, p_0, p_1)$, let $e_0(f)$ be the number of erroneous observations in those cases in which $f(x) = 0$, and $e_1(f)$ the number of errors when $f(x) = 1$, so that $e_0(f) + e_1(f) = e(f)$. Let $N_0(f)$ be the total number of observations made when $f(x) = 0$, and $N_1(f)$ the number when $f(x) = 1$. Then we have

$$Pr(Y \mid f, p_0, p_1) = (1 - p_0)^{N_0(f)-e_0(f)} p_0^{e_0(f)} (1 - p_1)^{N_1(f)-e_1(f)} p_1^{e_1(f)}. \quad (16)$$

For a fixed f, this likelihood is maximized by setting $p_0 = e_0(f)/N_0(f)$ and $p_1 = e_1(f)/N_1(f)$. An MLE is therefore obtained by maximizing

$$\left[1 - \frac{e_0(f)}{N_0(f)}\right]^{N_0(f)-e_0(f)} \left[\frac{e_0(f)}{N_0(f)}\right]^{e_0(f)} \left[1 - \frac{e_1(f)}{N_1(f)}\right]^{N_1(f)-e_1(f)} \left[\frac{e_1(f)}{N_1(f)}\right]^{e_1(f)} \quad (17)$$

over all $f \in F$.

When f is unrestricted, it is not difficult to see that an MLE can be obtained as before, using corollary 1. Both the error-minimizing and error-maximizing solutions are MLEs. However, when f is restricted, an MLE may be neither error minimizing nor error maximizing. The same is true *a fortiori* of bayesian estimates. Since this makes computation more difficult, we will assume hereafter that errors are symmetric.

3.5. CONFIDENCE LEVELS

Numerical regression analysis typically involves the computation of confidence intervals for estimates and tests for the statistical significance of the regression. Both have analogs in the boolean case.

In a numerical regression (1), a confidence interval for an estimate $\hat{g}(x)$ of a function value $g(x)$ is computed by determining the distribution of $\hat{g}(x)$ that would result if several samples were drawn from the original population. If the distribution is widely dispersed, the confidence interval is wide, and one has less confidence in the estimate. For example, if there is a 95% chance that $\hat{g}(x)$ deviates no more than $\pm\Delta$ from $g(x)$, then $\hat{g}(x) \pm \Delta$ is a "95% confidence interval". A similar analysis applies to parameter estimates.

A difficulty with this approach is that the distribution of $\hat{g}(x)$ depends on the unknown variance σ^2 of the error term ε. An estimate of σ^2 can be computed (perhaps as the mean squared error, adjusted for bias), but this estimate is itself a random variable whose value depends on the sample. The classical solution to this problem is to invent a "pivot", or a statistic that is a function of $g(x)$ and sample statistics but whose distribution is independent (or nearly independent) of σ^2. A 95% confidence interval for this statistic can then be transformed to a 95% confidence interval for $\hat{g}(x)$. The pivot traditionally used has a t distribution whose variance depends only on the sample size (or, more precisely, the number of "degrees of freedom") and is independent of σ^2.

In boolean regression, we can replace confidence intervals with confidence radii. That is, we can measure our confidence that the true function lies within a distance D of the estimated function $\hat{f}$. The distance between functions can be defined as the Hamming distance, which in this case is the number of arguments

$x \in X$ on which the functions differ. A 95% confidence radius D indicates that an estimate based on a random sample of size N has at least a 95% chance of lying within Hamming distance D of $\hat{f}$. The *confidence level* of $\hat{f}$ is the probability that a random sample yields $\hat{f}$ exactly.

Again, we have the difficulty that the distribution of $\hat{f}$ depends on the probability p of observational error, and in this case no "pivot" seems to be available. A further difficulty is that an adequate regression sometimes seems to *require* that we take account of the prior distribution of p, as discussed in the last section.

A natural solution to these difficulties is to use a bayesian approach to confidence testing. In this approach, the confidence level of an estimate $\hat{f}$ is simply the posterior probability that $\hat{f}$ is the true function. We have a 95% confidence radius D when the posterior probabilities of all functions f within Hamming distance D of $\hat{f}$ sum to at least 0.95.

The posterior probability of a function f is computed by integrating (11) with respect to p.

$$Pr(f\,|\,Y) = Pr(Y\,|\,f)Pr(f)/Pr(Y) = Pr(f)\int_0^1 Pr(Y\,|\,f, p)\pi(p)dp/Pr(Y). \qquad (18)$$

Thus, the confidence level of an estimate $\hat{f}$ is $Pr(\hat{f}\,|\,Y)$, and D is a 95% confidence radius if

$$\sum_{f \in F_D} Pr(f\,|\,Y) \geq 0.95,$$

where F_D is the set of all functions in F that lie within a Hamming distance D of $\hat{f}$. We will compute this confidence level for the two types of prior distributions of p discussed in the last section.

Model 1

p is uniformly distributed on $[0, \rho]$. Integrating by parts, $Pr(Y|f)$ in (18) becomes

$$Pr(Y\,|\,f) = \frac{1}{\rho}\int_0^\rho Pr(Y\,|\,f, p)dp$$

$$= \frac{\rho^{e(f)}}{(N+1)} \sum_{i=1}^{N-e(f)} c_i(1-\rho)^{N-e(f)-i}, \qquad (19)$$

where

$$c_i = \binom{N-e(f)}{i}\binom{N}{i}^{-1}. \qquad (20)$$

(When $\rho = 1$, we take the last term of the summation in (20) to be $c_{N-e(f)}$.) From (12) and (18), we have that the confidence level for an estimate f is

$$Pr(f \mid Y) = Pr(Y \mid f) \Big/ \sum_{f' \in F} Pr(Y \mid f'), \tag{21}$$

where $Pr(Y \mid f)$ is given by (20). The confidence level for the estimated value of f at a particular x is given by

$$Pr(f(x) \mid Y) = \sum_{f' \in F_x} Pr(Y \mid f') \Big/ \sum_{f' \in F} Pr(Y \mid f'), \tag{22}$$

where $F_x = \{ f' \in F \mid f'(x) = f(x) \}$.

In the example of table 1, the confidence levels for $\tilde{f} (= \hat{f})$ under Model 1 are 48.2%, 83.5%, 96.5% and 98.4%, respectively, when $\rho = 1$, 0.7, 0.5 and 0.3. The corresponding confidence levels for $\tilde{f}(0, 1, 0, 1, 1) = 1$ are higher because it is easier to get one value right than six: 50%, 86.6%, 99.9% and 99.99%. A 50% confidence level for $\rho = 1$ may seem counter-intuitive when $f(0, 1, 0, 1, 1) = 1$ in fifteen out of twenty-two observations, but recall that when nothing is known about the distribution of p, $\tilde{f}(0, 1, 0, 1, 1) = 0$ with $\tilde{p} = 15/22$ is an equally good explanation of the data.

Model 2

The probability density function of p is $\pi(p) = (1 + \kappa)(1 - p)^{\kappa}$. When κ is an integer, $Pr(Y \mid f)$ in (21) becomes

$$Pr(Y \mid f) = \frac{1}{(n+1)} \binom{N}{e(f) - \kappa}^{-1} \frac{e(f) + 1}{e(f) - \kappa + 1}. \tag{23}$$

The confidence level for an estimate f is again given by (21), and for a particular value $f(x)$ by (22), where $Pr(Y \mid f)$ is given by (23).

Under Model 2, the confidence levels in table 1 for $\tilde{f}$ are 48.2%, 87.2%, 96.7% and 98.4% for $\kappa = 0, 1, 2, 3$, and the corresponding levels for $\tilde{f}(0, 1, 0, 1, 1)$ are 50%, 89.5%, 98.6% and 99.8%.

3.6. SIGNIFICANCE OF REGRESSION

In traditional statistics, a regression is significant if the apparent relationship between the dependent and independent variables cannot be explained by chance. A maximum likelihood ratio test is used to check for significance, and the test is usually based on the fact that the ratio of two normalized sum-of-squares statistics has an F distribution. For the boolean case, we will again use a bayesian approach.

Suppose we wish to check whether the bayesian estimate $\tilde{f}$ of f captures a significant relationship between y and x in table 1. We can formulate two hypotheses:

H_0: $f(x)$ is the same for all $x \in X$ (i.e. x has no bearing on y);

H_1: $f(x)$ is not the same for all $x \in X$.

The question is whether we can reject the null hypothesis H_0 in favor of H_1. To answer it, we use an approach analogous to that of maximum likelihood ratio testing: we find the function $\tilde{f}_0$ having the highest posterior probability $Pr(\tilde{f}_0 | Y, H_0)$ under H_0, and the function $\tilde{f}$ having the highest posterior probability $Pr(\tilde{f} | Y, H_1)$ under H_1. The function $\tilde{f}_0$ is the bayesian estimate of f when F contains only the two constant functions given by $f(x) = 0$ for all $x \in X$ and $f(x) = 1$ for all $x \in X$. The function $\tilde{f}$ is of course the bayesian estimate of f using the original set F, which has already been computed. The significance level of the regression is the probability that $\tilde{f}_0$ is the true function, given that either $\tilde{f}_0$ or $\tilde{f}$ is the true function:

$$Pr(\tilde{f}_0 | Y, \tilde{f}_0 \text{ or } \tilde{f}) = \frac{Pr(\tilde{f}_0 | Y)}{Pr(\tilde{f}_0 | Y) + Pr(\tilde{f} | Y)}. \tag{24}$$

(Here we assume that $\tilde{f}_0 \neq \tilde{f}$.) The probabilities on the right are given by (21).

Let us analyze the regression for table 1 using Model 1's prior distribution for p, with $\rho = 1/2$. The MLE $\hat{f}_0$ under H_0 is clearly the constant function having the value 1, since it has forty-one errors, and the other constant function has forty-six. Since $41/87 \leq 1/2$, $\tilde{f}_0 = \hat{f}_0$. The posterior probability of $\tilde{f}_0$ is $Pr(\tilde{f}_0 | Y) = 0.0000128$. Since we found earlier that the confidence level for $\tilde{f}$ is $Pr(\tilde{f} | Y) = 0.9647$, the significance level of the regression is $0.0000128/(0.0000128 + 0.9647) = 0.0000132$. There is only a 0.00132% chance that the regression is not significant.

3.7. PRACTICAL CONSIDERATIONS

In practice, one would first compute the MLE of f by finding the error-minimizing and error-maximizing estimates and picking one that results in the higher likelihood $Pr(Y | f, p)$. If the resulting estimate is error minimizing, then it is probably adequate it stands. This is because in practically any imaginable application, a large error probability is, if anything, less likely than a small one. If the bayesian solutions were computed using a prior distribution that gives less weight to large error probabilities, the same error-minimizing estimate would ensue.

If the MLE is error maximizing, however, we must consider the prior distribution of the error probability p. We may use Model 1 or Model 2, but the latter may be more practical if $\hat{f}_3$ is difficult to calculate for Model 1. Under Model 2, we find the smallest κ for which the bayesian estimate is the error-minimizing function $\hat{f}_1$ rather than the error-maximizing function $\hat{f}_2$; that is, the smallest κ for which $e(\hat{f}_1) \leq N + \kappa - e(\hat{f}_2)$ We can then examine the prior distribution $\pi(p) = (1 + \kappa)(1 + p)^{\kappa}$ and judge whether large errors are as unlikely *a priori* as implied by this distribution. If so, we should use $\hat{f}_1$, and otherwise function $\hat{f}_2$.

Confidence levels and significance tests require that one choose a prior distribution of the error probability p, regardless of the outcome of the MLE calculation. If the MLE is error minimizing, however, one has a choice of either Models 1 or 2, whereas if it is error maximizing, Model 2 may be the only computationally practical option.

When it is difficult to say what the prior distribition of p is, it is good practice to obtain a lower bound on the confidence level by making a conservative assumption regarding the distribution. In Model 1, where $p \leq \rho$, a conservative interpretation would pick a fairly large value for ρ, perhaps 1/2. If one could derive a reasonably high confidence level on the very conservative assumption that the probability of error is less than 1/2, then the confidence level is itself conservatively estimated and probably much higher. In Model 2, a conservative interpretation would pick a small value for κ. If a reasonably high confidence level results from assuming $\kappa = 1$, for instance, than one has confidence in the regression. If an error-maximizing bayesian estimate of f is used, the κ used in measuring confidence must not be so large as to result in an error-minimizing bayesian estimate of f.

Similar considerations apply to the significance tests.

As for computing the error-minimizing and error-maximizing functions, this can in principle be done simply by enumerating all boolean functions in the given class F. But in special cases, such as those described in the remaining sections, there are more efficient methods. Computation of exact confidence and significance levels is difficult even in the special cases, but we will show how to obtain good estimates using a modest amount of computation.

4. Unrestricted functions

4.1. BAYESIAN ESTIMATION

The simplest special case is that in which F contains *all* boolean functions defined on X. That is, no restrictions are placed on the best-fitting boolean function. Although this case is of limited applicability, it provides a simple context in which to introduce methods for approximating confidence and significance levels. We later adapt them to more interesting cases.

Corollary 1 summarizes the obvious way to find error-maximizing and error-minimizing functions in this case. If there is no prior knowledge regarding the distribution of p, both are bayesian estimates of f, whereas if larger error probabilities p are less likely, the error-minimizing function is a bayesian estimate of f.

4.2. CONFIDENCE LEVELS AND SIGNIFICANCE OF REGRESSION

Computing exact confidence levels is difficult, since the marginal probability $Pr(Y)$ in (18) is given by a sum in (12) over all $f \in F$. However, it is relatively easy to obtain a close approximation to $Pr(Y)$.

We can obtain a lower bound on $Pr(Y)$ that approximates $Pr(Y)$ simply by summing over a subset of F in (12). One reasonable subset to use is the set F_D of all functions in F that lie within a given Hamming distance D of the best fit $\tilde{f}$. Most of the probability $Pr(Y)$ is contributed by functions within a small Hamming distance from $\tilde{f}$, since they tend to result in fewer errors than those more distant. Thus, we have an approximation P_0 of $Pr(Y)$:

$$P_0 = \sum_{f' \in F_D} Pr(Y \mid f'). \tag{25}$$

Using (21), we can estimate the confidence level $Pr(f \mid Y)$:

$$Pr(f \mid Y) \approx Pr(Y \mid f)/P_0. \tag{26}$$

As an example, we will compute a confidence level for $\tilde{f}$ in the example of table 1. We found earlier that, using probability Model 1 with $\rho = 0.5$, the exact value of the confidence level (to four decimal places) is $P(\tilde{f} \mid Y) = 0.9647$. This computation required the summation of $2^6 = 64$ terms in (12). If we sum over only the seven terms in F_1 (i.e. all functions within a Hamming distance $D = 1$ of $\tilde{f}$), we obtain a close approximation, $P_0 = 0.9608$. If we sum over the twenty-two terms in F_2, $P_0 = 0.9643$.

Once confidence levels are computed, it is easy to obtain significance levels using (24).

4.3. EXTENSION TO A FULLY DEFINED FUNCTION

The regression described above defines $f(x)$ for observed values of x. If this partially defined boolean function is extended to an unrestricted fully defined function $f^*, f^*(x)$ is of course completely arbitrary for every unobserved x. It may be useful, however, to find a relatively simple logical formula that agrees with f on all observed x's. One may wish the formula to be a disjunction of as few terms as possible, or to involve as few x_j's as possible. Techniques for obtaining such a formula are discussed in detail in [5].

5. Positive functions

5.1. BAYESIAN ESTIMATION

It is frequently the case that one has to find a good boolean regression in the presence of some *a priori* structural information on the function f, e.g. f must be a monotone boolean function, or more general, there is known a partial order $\prec$ on the set of $0-1$ vectors x and f must satisfy the condition that $f(x) \leq f(x')$ whenever $x \prec x'$. Such boolean functions will be called *$\prec$-monotone.* (A more general notion of monotonicity and the corresponding "best fit" problem have been studied in [3].)

In this section, we consider the problem of finding the error-minimizing $\prec$-monotone boolean function for the given data, which consists of a set X of $0-1$ vectors and integers $t^0(x)$ and $t^1(x)$ for every $x \in X$, denoting the multiplicities of the false and true outcomes when x is observed. The number of errors is given by

$$e(f) = \sum_{x \in X: f(x)=1} t^0(x) + \sum_{x \in X: f(x)=0} t^1(x). \tag{27}$$

We can formulate the regression problem considered in this section as follows:

$$\begin{aligned} &\text{minimize} && e(f) \\ &\text{subject to} && f : X \to \{0,1\}, \quad \text{and} \\ & && f(x) \le f(x') \quad \text{for all } x, x' \in X \text{ with } x \prec x'. \end{aligned} \tag{28}$$

We shall show in the sequel that this problem can be solved in polynomial time, in at most $O(|X|^3)$ steps via a minimum cut computation in an associated capacitated network.

It is relatively easy to transform the above problem into a "maximum closure" problem, which then can be solved as a maximum flow problem (see e.g. [14]), and which thus provides a polynomial time solution to (28). However, for the sake of completeness, we give here another approach which is slightly different from the one in [14]. The key to this approach is to show that one can always find a minimum-capacity *monotone cut* in a flow network by modifying the arc capacities and solving an ordinary minimum cut problem on the modified network.

Let $G = (N, A)$ be an acyclic directed graph, and let c_{ij} be nonnegative capacities associated with the arcs $(i, j) \in A$. Two distinguished vertices of the network, $s \in N$ and $t \in N$, will be called the source and the sink, respectively. An *(s, t)-cut* is defined as the set of arcs $\{(i, j) \in A \,|\, i \in S, j \notin S\}$ connecting the vertices of a subset S to the vertices not in S, and where we assume that $s \in S$ and $t \notin S$. Such a cut will be denoted by C_S, referring to the fact that it is induced by the subset S. The capacity $c(C)$ of a cut C is defined as the sum of the capacities of the arcs in the cut, i.e. $c(C) = \sum_{(i,j) \in C} c_{ij}$.

The *minimum cut* problem is to find the subset S such that $s \in S$, $t \notin S$ and $c(S)$ is as small as possible. It is well known that this problem can be solved in $O(|N|^3)$ time, see e.g. [11].

A cut C_S is called *monotone* if there are not arcs directed backwards, i.e. there is no arc $(i, j) \in A$ such that $i \notin S$ and $j \in S$.

The *minimum monotone cut* problem is to find a subset S of the nodes such that $s \in S$, $t \notin S$, C_S is monotone and $c(C_S)$ is as small as possible. One way to find a minimum monotone cut for any acyclic directed path G is to find a minimum cut for a graph G' constructed on the same node set as follows. Let G' contain all the arcs of G with the same capacities. In addition, for every arc (i, j) of G, add to G' the reverse arc (j, i) with the infinite capacity. A minimum cut for G' will not contain any of the infinite-capacity arcs and will therefore be a minimum monotone cut for G.

We will show, however, how to solve the minimum monotone cut problem in a way that is more efficient because it does not increase the size of the network.

THEOREM 2

Let $G = (N, A)$ be an arbitrary acyclic directed graph with nonnegative real capacities c_{ij} on the arcs $(i, j) \in A$, and let M be an integer, $M > \frac{1}{2} |N| \sum_{(i,j) \in A} c_{ij}$. Then there are capacities $\tilde{c}_{ij}$, $(i, j) \in A$, such that the minimum cut C_S in the network $(N, A, \tilde{c})$ will be the minimum monotone cut in the original network (N, A, c), and such that

$$\tilde{c}(C_S) = M|N| + c(C_S). \tag{29}$$

Proof

For $i \in N$, let d_i^+ (d_i^-) denote the out-degree (in-degree) of vertex i, i.e. the number of edges going out from (entering) vertex i. For an arc $(i, j) \in A$, let P_{ij}^+ denote a shortest path in the network from j to t, and similarly, let P_{ij}^- denote a shortest path in the network from s to i.

Finally, let

$$\tilde{c}_{ij} = c_{ij} + M\left(\frac{1}{d_i^+} + \frac{1}{d_j^-} + \sum_{(k,l): P_{kl}^+ \ni (i,j)} \frac{1}{d_k^+} + \sum_{(k,l): P_{kl}^- \ni (i,j)} \frac{1}{d_l^-}\right). \tag{30}$$

In other words, the capacity of every arc on the path $\{(i, j)\} \cup P_{ij}^+$ is increased by M/d_i^+ and the capacity of every arc on the path $\{(i, j)\} \cup P_{ij}^-$ is increased by M/d_j^-, consecutively, for all arcs $(i, j) \in A$.

Now for a subset S of the vertices with $s \in S$, $t \notin S$ we have that C_S must contain at least one arc of the path $\{(i, j)\} \cup P_{ij}^+$ for all arcs $(i, j) \in A$ leaving i if $i \in S$, and for $j \notin S$ the cut C_S must contain at least one arc from $\{(i, j)\} \cup P_{ij}^-$ for all arcs $(i, j) \in A$ entering j. Therefore,

$$\tilde{c}(C_S) \geq c(C_S) + \sum_{i \in S} \sum_{j:(i,j) \in A} \frac{M}{d_i^+} + \sum_{j \notin S} \sum_{i:(i,j) \in A} \frac{M}{d_j^-} = c(C_S) + M|N|, \tag{31}$$

with equality for all monotone cuts.

If C_S is not monotone, i.e. if there is an arc $(k, l) \in A$ such that $k \notin S$ and $l \in S$, then, in addition to the above, C_S must contain at least one arc from the paths $\{(k, l)\} \cup P_{kl}^-$ and $\{(k, l)\} \cup P_{kl}^+$, which implies that

$$\begin{aligned} \tilde{c}(C_S) &\geq c(C_S) + M|N| + \frac{M}{d_l^-} + \frac{M}{d_k^+} \\ &\geq c(C_S) + M|N| + \frac{2M}{|N|} > M|N| + \sum_{(i,j) \in A} c_{ij}. \end{aligned} \tag{32}$$

This means that

$$\tilde{c}(C_{S'}) > \tilde{c}(C_S) \tag{33}$$

whenever $C_{S'}$ is non-monotone and C_S is monotone. Since there are monotone cuts in the network, the minimum cut must be monotone by (33) and hence proving the theorem by (31).

It is easy to implement the computations of $\tilde{c}$ such that all arc capacities can be updated in at most $O(|N|\ |A|)$ steps, thus proving that a minimum monotone cut can be computed in at most $O(|N|^3)$ time. □

Let us now return to the best-fit problem (28). We shall build a network, associated with the data (X, t^0, t^1), such that there will be a one-to-one correspondence between the monotone cuts C_S and the $\prec$-monotone mappings $f: X \to \{0, 1\}$. Moreover, the capacity $c(C_S)$ of the cut and the error $e(f)$ will be equal. Thus, the computation of a minimum monotone cut in that network will give us the optimal solution of problem (28).

Let $N = X \cup \{s, t\}$ be the set of vertices, and put an arc from s to every element of X, from every element of X to t, and from x to x' whenever $x \prec x'$. For the sake of simplicity, we can delete all arcs (i, j) for which there is another vertex k such that (i, k) and (k, j) are both arcs. Let A denote the set of arcs finally left.

Now, if $f: X \to \{0, 1\}$ is a monotone mapping, i.e. there is no $x \prec x'$ such that $f(x) = 1$ and $f(x') = 0$, then $S_f = \{s\} \cup \{x \in X | f(x) = 0\}$ induces a monotone cut in (N, A). Conversely, if C_S is a monotone cut for a subset $S \subset N$ such that $s \in S$ and $t \notin S$, then $f_S(x) = 1$ iff $x \notin S$ for $x \in X$ defines a $\prec$-monotone mapping.

For every $x \in X$, let P_x^+ denote a shortest path from x to t, and let P_x^-denote a shortest path from s to x in the network (N, A). Moreover, let the capacities c_{ij} be defined by

$$c_{ij} = \sum_{P_x^+ \ni (i,j)} t^1(x) + \sum_{P_x^- \ni (i,j)} t^0(x). \tag{34}$$

Since a monotone cut C_S intersects every path P_x^+ for $x \in S$ and P_x^- for $x \notin S$ in exactly one arc, we have

$$c(C_S) = \sum_{x \in S} t^1(x) + \sum_{x \notin S} t^0(x) = e(f_S). \tag{35}$$

The complexity of building up the network and computing the capacities is of $O(|X|^3)$, and thus the total complexity of solving problem (28) is also $O(|X|^3)$.

5.2. EXTENSION TO A FULLY DEFINED FUNCTION

A fully defined function f^* to which $\tilde{f}$ is extended must satisfy

$$\max_{x' \prec x} \tilde{f}(x') \leq f^*(x) \leq \min_{x' \succ x} \tilde{f}(x')$$

for all boolean vectors x, $f^*(x)$ is determined for an unobserved x when these upper and lower bounds are equal.

5.3. CONFIDENCE LEVELS AND SIGNIFICANCE OF REGRESSION

Confidence and significance levels can be computed as in section 4.2, except that F_D in (25) becomes the set of all *monotone* (or $\prec$-monotone) boolean functions defined on X that are within a Hamming distance D of the best fit $\tilde{f}$. For small D, these function can be enumerated quickly if one does not change $\tilde{f}(x)$ to 1 without first changing $\tilde{f}$ for all immediate successors of x for which $\tilde{f}$ is 0, and one does not change $\tilde{f}(x)$ to 0 without first changing $\tilde{f}$ for all immediate predecessors of x for which $\tilde{f}$ is 1.

6. Functions defined by logical formulas

6.1. BAYESIAN ESTIMATION

We now consider the problem of estimating a regression formula expressed in propositional logic. We will show that the bayesian estimation problem can be solved as a pseudo-boolean optimization problem.

Suppose that we wish to find a regression formula in the form of (3) to fit the data in table 1. In this case, (3) becomes

$$\beta_0 \vee \beta_1 x_1 \vee \ldots \vee \beta_5 x_5, \tag{36}$$

where each $\beta_i \in \{0, 1\}$. Recall that we can find a bayesian estimate if we can find an error-minimizing and an error-maximizing function in F (and perhaps an error-maximizing function subject to $\tilde{p} \leq \hat{p}$). In the present case, F contains all fully defined functions that are expressed by a formula in the form of (36).

We will minimize and maximize error $e(f) = e(\beta)$ by writing $e(f)$ as a pseudo-boolean (i.e. real-valued) function of the boolean arguments $\beta_0, \ldots, \beta_5$, and then minimizing or maximizing the function.

To see the principle involved, let us first count the errors resulting from the first line of table 1. If $\beta_0 = \beta_2 = \beta_4 = \beta_5 = 0$, it is clear that $f_\beta(0, 1, 0, 1, 1) = 0$, resulting in seven erroneous observations. Otherwise, $f_\beta(0, 1, 0, 1, 1) = 1$, and we have fifteen errors. So, the number of errors resulting from line 1 is

$$7(1 - \overline{\beta}_0 \overline{\beta}_2 \overline{\beta}_4 \overline{\beta}_5) + 15 \overline{\beta}_0 \overline{\beta}_2 \overline{\beta}_4 \overline{\beta}_5,$$

where $\overline{\beta}_i = 1 - \beta_i$. Summing similar expressions for all six lines of table 1 and collecting terms, we obtain the pseudo-boolean function

$$\begin{aligned} e(\beta) = 41 &+ 8\overline{\beta}_0 \overline{\beta}_2 \overline{\beta}_4 \overline{\beta}_5 - 6\overline{\beta}_0 \overline{\beta}_1 \overline{\beta}_2 \overline{\beta}_5 - 6\overline{\beta}_0 \overline{\beta}_1 \overline{\beta}_4 \overline{\beta}_5 \\ &+ 4\overline{\beta}_0 \overline{\beta}_1 \overline{\beta}_5 + 12\overline{\beta}_0 \overline{\beta}_1 \overline{\beta}_2 \overline{\beta}_4 - 7\overline{\beta}_0 \overline{\beta}_3 \overline{\beta}_4 \overline{\beta}_5. \end{aligned} \tag{37}$$

The problem of solving a pseudo-boolean optimization problem such as this one is well studied [9,8]. In this case, the minimum error solution is $(\beta_0,\dots,\beta_5) = (0, 0, 1, 0, 0, 0)$, with $e(\beta) = 32$ errors. This corresponds to the propositional function

$$f(x) = x_2. \tag{38}$$

The maximum error solutions are $\beta = (0, 0, 0, 1, 0, 1)$ and $(0, 0, 0, 1, 0, 0)$ and $(0, 0, 0, 0, 0, 1)$, all with $e(\beta) = 53$ errors, corresponding to the propositional functions

$$f(x) = x_3 \vee x_5, \tag{39}$$

$$f(x) = x_3, \tag{40}$$

$$f(x) = x_5. \tag{41}$$

Since $32 < N - 53(=87 - 53 = 34)$, theorem 1 implies that (38) is the estimate we want.

Suppose in general that the regression formula is disjunctive:

$$\bigvee_{i=1}^{m} \beta_i t_i(x), \tag{42}$$

where each $t_i(x)$ is a boolean function of x that involves no β_i's. For instance, $t_i(x)$ may be a conjunction of literals (terms of the form x_j or $\bar{x}_j$), in which case (42) is in disjunctive normal form. (Any propositional formula is equivalent to some formula in this form.) Let $t^0(x)$ by the number of observations of $f(x)$ for which $y = 0$, and $t^1(x)$ the number for which $y = 1$. Then the error can be expressed as

$$e(\beta) = \sum_{x \in X} t^0(x) + \sum_{x \in X} (t^1(x) - t^0(x)) \prod_{k \in K(x)} \bar{\beta}_k, \tag{43}$$

where $K(x) = \{k \mid t_k(x) = 1\}$.

It may be useful to let the regression formula be conjunctive,

$$\bigwedge_{k=1}^{m} t_i(x)^{\beta_i}, \tag{44}$$

where each $t_i(x)$ appears if $\beta_i = 1$ and is absent if $\beta_i = 0$. If $t_i(x)$ is a disjunction of literals, (44) is in conjunctive normal form (into which any propositional formula can be put). The error is

$$e(\beta) = \sum_{x \in X} t^1(x) + \sum_{x \in X} (t^0(x) - t^1(x)) \prod_{k \in K(x)} \beta_j. \tag{45}$$

6.2. CONFIDENCE LEVELS

The confidence level $Pr(f_{\bar{\beta}} \mid Y)$ for an estimate $f_{\bar{\beta}}$ is given by (21), where F is the set of all boolean functions f_β expressed by formulas having the desired form.

In the example of table 1, the confidence level of the regression formula (38) is 28.2%, assuming the prior distribution of Model 1 with $\rho = 0.5$. This suggests that a more complex model is needed, perhaps one with negated variables $\beta_i \bar{x}_i$. Another option is to use "associations" (to borrow a term from categorical data analysis), which are quadratic terms $\beta_{ij} x_i x_j$. Or we may even want to use "interactions", or cubic terms $\beta_{ijk} x_i x_j x_k$.

Confidence radii can be defined in the context of formulas by using the Hamming distance between coefficient vectors β. In the example, we can have 43.3% confidence that the true β is within a radius of one of $\tilde{\beta}$. This means that there is a probability of 0.433 that the true function differs from (38) in at most one coefficient. A radius of 2 yields only a 57.2% confidence level.

6.3. SIGNIFICANCE OF REGRESSION

To compute significance of regression in the example of table 1, we consider the hypotheses

H_0: $\beta_i = 0$ for $i = 1,\ldots,5$ (i.e. x has no bearing on y);

H_1: $\beta_i \neq 0$ for at least one $i \in \{1,\ldots,5\}$.

The best fit $\tilde{f}_0$ under H_0 is defined by $\beta_0 = 1$ (with all other $\beta_i = 0$), since we saw earlier that $f(x) = 1$ is the better fitting constant function. The best fit $\tilde{f}$ under H_1 is defined by (38). After using (21) to obtain a confidence level $Pr(\tilde{f}_0 \,|\, Y) = 0.0104$ for $\tilde{f}_0$, we have from (24) that the regression is significant at level 0.036.

Since we can have only 28% confidence in the regression formula (38), we should consider adding some additional terms to (36). In numerical regression analysis, it is common to use a stepwise procedure whereby one adds new groups of variables to the regression formula, one group at a time, and each time checks for significant improvement in the fit. An analogous procedure is available in boolean regression.

For instance, we may want to add some negated variables to (36) to obtain a function that need not be positive:

$$\beta_0 \vee \beta_1 x_1 \vee \ldots \vee \beta_5 x_5 \vee \gamma_1 \bar{x}_1 \vee \ldots \vee \gamma_5 \bar{x}_5 . \tag{46}$$

The best fit $\tilde{f}$ having the form (46) is

$$f(x) = x_2 \vee \bar{x}_4 . \tag{47}$$

The corresponding confidence level is only $Pr(\tilde{f} \,|\, Y) = 11.2\%$. This time, the competing hypotheses are

H_0: $\beta_i = 0$ for $i = 6,\ldots,10$ (i.e. the negated terms have no bearing on y);

H_1: $\beta_i \neq 0$ for at least one $i \in \{6,\ldots,10\}$.

If we cannot reject H_0, then the fit is not significantly improved.

The best fit $\tilde{f}_0$ for H_0 is the best fit when only the positive terms are used, namely (38). We must recompute its confidence level, however, since the set F now contains all functions having the form (46). We obtain $Pr(\tilde{f}_0 | Y) = 0.00878$. The best fit under H_1 is of course (46). Using (24), we compute that we can reject H_0 with only a 0.073 significance level. Not only do we have little confidence (11%) in the new formula, but we cannot state that it is a better fit than (38) with, say, a 5% level of significance.

References

[1] A. Agresti, *Categorical Data Analysis* (Wiley, New York, 1990).

[2] E.B. Andersen, *The Statistical Analysis of Categorical Data* (Springer, Berlin/New York, 1990).

[3] E. Boros, P.L. Hammer and J.N. Hooker, Predicting cause–effect relationships from incomplete discrete observations, SIAM J. Discr. Math. 7(1994)423–435.

[4] L. Breiman, *Classification and Regression Trees* (Wadsworth, Belmont, CA, 1984).

[5] Y. Crama, P.L. Hammer and T. Ibaraki, Cause–effect relationships and partially defined Boolean functions, Ann. Oper. Res. 16(1988)299–325.

[6] D. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison–Wesley, Reading, MA, 1989).

[7] L.A. Goodman and C.C. Clogg, *The Analysis of Cross-classified Data Having Ordered Categories* (Harvard University Press, Cambridge, MA, 1984).

[8] P. Hammer and S. Rudeanu, *Boolean Methods in Operations Research and Related Areas* (Springer, Berlin, 1968).

[9] P. Hansen and B. Jaumard, Algorithms for the maximum satisfiability problem, Computing 44(1990)279–303.

[10] J.H. Holland, K.J. Holyoak, R.E. Nisbett and P.R. Thagard, *Induction: Process of Inference, Learning and Discovery* (MIT Press, Cambridge, MA, 1989).

[11] A.V. Karzanov, Determining the maximal flow in a network by the method of preflows, Sov. Math. Doklady 15(1984)434–437.

[12] R. Michalski, J. Carbonell and T. Mitchell (eds.), *Machine Learning: An Artificial Intelligence Approach* (Tioga Press, Palo Alto, CA, 1983).

[13] N.J. Nilsson, *The Mathematical Foundations of Learning Machines* (Morgan Kaufmann, San Mateo, CA, 1990).

[14] J.-C. Picard, Maximal closure of a graph and applications to combinatorial problems, Manag. Sci. 22(1976)1268–1272.

[15] S.J. Press, *Bayesian Statistics: Principles, Models and Applications* (Wiley, New York, 1989).