Annales Univ. Sci. Budapest., Sect. Comp. 13 (1992) 179-194

THE PROBLEM OF LEARNING CONCEPTS. A PROBABILISTIC VIEW

E. Alvarez (Santander, Cantabria, Spain)
A. Benczúr (Budapest, Hungary)
E. Castillo (Santander, Cantabria, Spain)
J.M. Sarabia (Santander, Cantabria, Spain)

Abstract. The problem of learning some kind of probabilistic Boolean concepts, which are an extension of the Boolean concept used by Valiant, is analyzed. An algorithm based on the maximum likelihood principle is given for learning these concepts from neutral examples by means of multinomial and Poissonian schemes. Asymptotic results, based on the delta method, allow the characterization of classes of learnable and non-learnable concepts. Finally, two illustrative examples of application are given.

1. Introduction

The problem of learning Boolean concepts has been investigated by many authors as Angluin (1978, 1986), Dietterich and (1983), (1984), and (1988), etc. They use a deterministic view of concepts which is useful when no uncertainty associated with them exists. However, there are many practical problems, as medical diagnosis for example, in which a different view of concepts is required.

In this paper we analyze the problem of learning concepts from a probabilistic point of view and show its practical interest by means of very simple practical examples.

2. Probabilistic view of learning

In the theory of learning (deterministic) Boolean concepts Valiant [3], Pitt and Valiant [2], we assume that each object in our world is represented by some assignment of the feature variables $\{x_i\}$ to either 0 or 1.

Thus, each object is simply a vector $\vec{x} \in \Omega = \{0, 1\}^t$. A concept C is a subset of the 2^t possible vectors. But in many practical problems there is an inherent uncertainty: the observation of the feature variables does not fully determine whether the object belongs to the concept or not. Let us suppose that it is because of the

existence of another set of unobservable feature variables $\{Y_i\}$, $i = 1, 2, \ldots, k$. So, each object is represented by the concatenated vector $(\vec{x}, \vec{y}) \in \{0, 1\}^{t+k}$, but only $\vec{X}$ is observable, and the full description of a concept is given by $D^+ \subset \{0, 1\}^{t+k}$.

Following Valiant's approach, we assume that examples are generated by a fixed, but unknown probability distribution $\pi(\vec{x}, \vec{y})$ on $\{0, 1\}^{t+k}$. The marginal distribution of $\vec{X} \in \{0, 1\}^t$ is denoted by $m(\vec{x}) = \sum_{\vec{y} \in \{0, 1\}^k} \pi(\vec{x}, \vec{y})$, and the conditional probability that the object belongs to the concept after observing $\vec{x}$ is $p(\vec{x}) = \frac{\sum_{(\vec{x}, \vec{y}) \in D^+} \pi(\vec{x}, \vec{y})}{m(\vec{x})}$.

In the following we shall assume that the set Ω is partitioned in three sets

$$\begin{aligned} \Omega_0 &= \{x \in \Omega / p(x) = 0 \text{ or } p(x) = 1\}, \\ \Omega_1 &= \{x \in \Omega / 0 < p(x) < 1\}, \\ \Omega_2 &= \{x \in \Omega / p(x) \text{ known}\}. \end{aligned}$$

Note that the deterministic concepts correspond to the special case where $\Omega_1 = \Omega$.

Now we can state that the only goal of learning a probabilistic concept can be the estimation somehow of the distribution $p(\vec{x})$. Thus, we assume that m is only a parameter of the observation process during the learning and after it as well.

The formal description of the learning process is the following: Let P be the probabilistic concept to be learnt. P is given by one function $\left\{p(\vec{x}), \vec{x} \in \Omega = \right.$ $\left. = \{0, 1\}^t\right\}$, where $p(\vec{x})$ represents the probability of an object with associated vector $\vec{x}$ to be a positive example of P.

During the learning process, an observation consists of a completely or incompletely specified feature vector $\vec{x}^* \in \Omega^* = \{0, 1, *\}^t$ and a value $h =$ " + " if the observed example belongs to the concept and $h =$ " − " if it does not. Note that each $\vec{x}^* \in \Omega^*$ represents a subset $C_{\vec{x}^*} = \{\vec{x} \in \Omega / x_i = x_i^* \text{ if } x_i^* \neq "*", \forall i\}$. A vector $\vec{x}^* \in \Omega^*$ is called total if every variable is determined, i.e. if $\vec{x}^* \in \Omega$, otherwise it is called a partial vector. Functions m and p can be easily extended to Ω^*:

$$m(\vec{x}^*) = \sum_{\vec{x} \in C_{\vec{x}^*}} m(\vec{x}); \quad \forall \vec{x}^* \in \Omega^* \tag{1}$$

and

$$p(\vec{x}^*) = \frac{\sum_{\vec{x} \in C_{\vec{x}^*}} m(\vec{x}) p(\vec{x})}{m(\vec{x}^*)}; \quad \forall\, \vec{x}^* \in \Omega^* \tag{2}$$

During the learning process, observations are obtained. After a number of observations n, or after a time τ having a random number of observations n_τ, the learning process stops and deduces a program that computes a function $q(\vec{x})$ for every $\vec{x} \in \Omega$. The function $q(\vec{x})$ (an estimate of $p(x)$) is the learnt concept, and we can use it only to guess the probability that a real object with feature vector $\vec{x}$ belongs to the concept.

The goodness of $q(\vec{x})$ is measured by the expected value of a loss function $r(p(\vec{x}), q(\vec{x}))$, and it is given by

$$L(q) = E\left[r\left(p(\vec{x}), q(\vec{x})\right)\right] = \sum_{\vec{x} \in \Omega} m(\vec{x})\, r\left(p(\vec{x}), q(\vec{x})\right). \tag{3}$$

We say the accuracy of $q(\vec{x})$ is ε if

$$\sum_{\vec{x} \in \Omega} m(\vec{x})\, r\left(p(\vec{x}), q(\vec{x})\right) \leq \varepsilon. \tag{4}$$

Since the observation process is random, we cannot assure an accuracy ε with probability 1 for arbitrarily small ε. We can only give a level α, and assure that

$$\Pr\left(\sum_{\vec{x} \in \Omega} m(\vec{x})\, r\left(p(\vec{x}), q(\vec{x})\right) \leq \varepsilon\right) \geq 1 - \alpha, \tag{5}$$

where Pr means the probability distribution of the sampling process.

Some important loss functions are:

(i) Bounded square:

$$r_1(z) = \begin{cases} 0, & \text{if } z^2 \leq \delta^2 \\ 1, & \text{if } z^2 > \delta^2 \end{cases}, \quad \text{then}$$

$$L_1(q) = \sum_{(p(\vec{x}) - q(\vec{x}))^2 > \delta^2} m(\vec{x}). \tag{6}$$

(ii) Quadratic loss:

$$r_2 = z^2, \quad \text{then}$$

$$L_2(q) = \sum_{\vec{x}\in\Omega} m(\vec{x}) \left(p(\vec{x}) - q(\vec{x})\right)^2. \tag{7}$$

(iii) In the case we have to make a decision after observing the feature variables $\vec{x}$ of an object that it belongs to the concept or not, and we will lose 1 if the decision is wrong, the optimal decision based on $q(\vec{x})$ is

$$d(q(\vec{x})) = \begin{cases} "+" & \text{if } q(\vec{x}) \geq \frac{1}{2}, \\ "-" & \text{if } q(\vec{x}) < \frac{1}{2}. \end{cases} \tag{8}$$

The corresponding loss function has two variables, and is:

$$r_3(z,v) = \begin{cases} z & \text{if } v < \frac{1}{2}, \\ \frac{1}{2} & \text{if } v = \frac{1}{2}, \\ 1-z & \text{if } v > \frac{1}{2}; \end{cases} \quad \text{then,} \tag{9}$$

$$L_3(q) = \sum_{q(\vec{x})<\frac{1}{2}} m(\vec{x})\, p(\vec{x}) + \sum_{q(\vec{x})>\frac{1}{2}} m(\vec{x})[1 - p(\vec{x})] + \frac{1}{2} \sum_{q(\vec{x})=\frac{1}{2}} m(\vec{x}).$$

The main problem in the theory of learnable (see Valiant [3]) is to find classes of concepts that are learnable in feasible time. We extend Vailant's definition of learnability for probabilistic concepts in the following way:

Definition 1. Let $\mathcal{F}$ be a class of probabilistic concepts. Let r be a given loss function, with associated parameter $\delta < 0$, and $\varepsilon > 0$ and $\alpha < 0$ denote the accuracy and the level of learning. We say $\mathcal{F}$ is learnable from examples iff there exists a polynomial $G(u,v,w,z)$ and a learning algorithm A such that $\forall P \in \mathcal{F}$, with associate function p, the algorithm A halts in time, or after observing a number of examples, $G(T(P), \delta^{-1}, \varepsilon^{-1}, \alpha^{-1})$, where $T(P) = t$ is a measure of the size of P, and A outputs a function q, such that the concept Q, given by q belongs to $\mathcal{F}$, and

$$\Pr\left(\sum_{\vec{x}\in\Omega} m\Big(\vec{x}\ r(p(\vec{x})), q(\vec{x}), \delta\Big) \leq \varepsilon\right) \geq 1 - \alpha$$

for every distribution $m(\vec{x})$.

3. Learning algorithm

In this section we give two algorithms to learn a concept P, based on multinomial and Poissonian sampling.

3.1. Multinomial sampling

We suppose that during the learning process observations are independent, identically distributed random vectors, and that the probability of observing $\{\vec{x}, +\}$ and $\{\vec{x}, -\}$ are $m(\vec{x})p(\vec{x})$ and $m(\vec{x})[1 - p(\vec{x})]$, respectively. We also assume that we have at hand a sample of n neutral objects not necessarily total. In other words we can obtain the following set, given by the sample:

$$S = \left\{ n_x^h / x \in \Omega^* ; h \in \{+, -\} \right\}; \quad \sum_{n_x^h \in S} n_x^h = n, \tag{10}$$

where n_x^h is the number of objects in the sample of the type $\{x, h\}$.

Consequently, the likelihood function of the sample, which depends on the set of parameters $\{p(x)/x \in \Omega\}$, is given by

$$V = \prod_{y \in \Omega} \left\{ [p(y)]^{n_y^+} [1 - p(y)]^{n_y^-} \right\} \times$$

$$\times \prod_{y \in \Omega^* - \Omega} \left\{ \left(\sum_{z \in C_y} m(z)p(z) \right)^{n_y^+} \left(\sum_{z \in C_y} m(z)[1 - p(z)] \right)^{n_y^-} \right\} \tag{11}$$

and its logarithm becomes

$$L = \log V = \sum_{y \in \Omega} \{ n_y^+ \log p(y) + n_y^- \log[1 - p(y)] \} +$$

$$+ \sum_{y \in \Omega^* - \Omega} \left\{ n_y^+ \log \left[\sum_{z \in C_y} m(z)p(z) \right] + n_y^- \log \left[\sum_{z \in C_y} m(z)(1 - p(z)) \right] \right\}. \tag{12}$$

By derivation we get the likelihood equations:

$$\frac{\partial L(q)}{\partial p(x)} = \frac{n_x^+}{q(x)} - \frac{n_x^-}{1-q(x)} +$$

$$+ \sum_{y \in S_x} \left[\frac{n_y^+ m(x)}{\sum_{z \in C_y} m(z) q(z)} - \frac{n_y^- m(x)}{\sum_{z \in C_y} m(z)[1-q(z)]} \right] = 0 ; \quad x \in \Omega, \tag{13}$$

where

$$S_{\vec{x}} = \{ \vec{x}^* \in \Omega^* - \Omega / \ \vec{x} \in C_{\vec{x}^*} \} \tag{14}$$

and $q(x)$ is the maximum-likelihood estimate of $p(x)$.

3.2. Poissonian sampling

Now we assume a Poissonian sampling of duration τ, i.e. that the random variables n_x^+ and n_x^- are independent Poisson random variables with mean values $m(x)p(x)\tau$ and $m(x)[1-p(x)]\tau$, respectively. We also assume that we can obtain the following set, given by the sample:

$$S = \left\{ n_x^h / x \in \Omega^* ; \ h \in \{+,-\} \right\} \tag{15}$$

where, as before, n_x^h is the number of objects in the sample of the type $\{x, h\}$.

The likelihood function of the sample is

$$V = \prod_{y \in \Omega} \left\{ \frac{\exp\{-\tau m(y)\} [m^+(y)]^{n_y^+} [m^-(y)]^{n_y^-} (\tau)^{n_y^+ + n_y^-}}{n_y^+ n_y^-} \right\} \times \tag{16}$$

$$\times \prod_{y \in \Omega^* - \Omega} \left\{ \frac{\exp\left\{ -\tau \sum_{z \in C_y} m(z) \right\} \left[\sum_{z \in C_y} m^+(z) \right]^{n_y^+} \left[\sum_{z \in C_y} m^-(z) \right]^{n_y^-} (\tau)^{n_y^+ + n_y^-}}{n_y^+ n_y^-} \right\},$$

where

$$m^+(y) = m(y)p(y); \quad m^-(y) = m(y)[1-p(y)]. \tag{17}$$

Eliminating constant terms in (16) we get (11). Thus, $q(x)$ estimates coincide in both sampling schemes.

3.3. Estimates for total sample vectors

If all vectors in the sample are total the system (13) becomes

$$\frac{n_x^+}{q(x)} - \frac{n_x^-}{1-q(x)} = 0\,;\quad x \in \Omega \Longrightarrow q(x) = \frac{n_x^+}{n_x^+ + n_x^-}. \tag{18}$$

Note that $q(x)$ is well defined unless $n_x^+ + n_x^- = 0$. If this happens we shall do $q(x) = 1/2$. Thus, we have

$$q(x) = \begin{cases} \dfrac{n_x^+}{n_x^+ + n_x^-} & \text{if } n_x^+ + n_x^- > 0, \\ & \text{if } x \notin \Omega_2; \\ \dfrac{1}{2} & \text{if } n_x^+ + n_x^- = 0, \\ p(x) & \text{if } x \in \Omega_2. \end{cases} \tag{19}$$

The probabilities associated with the two different cases in the above expression where $x \notin \Omega_2$ are $1 - [1 - m(x)]^n$ and $[1 - m(x)]^n$, respectively, if we have multinomial sampling and $1 - \exp[-\tau m(x)]$ and $\exp[-\tau m(x)]$, respectively, if the sampling scheme is Poissonian. Thus, we can handle the random variable $q(x)$ as a linear convex combination of two obvious variables with the above probabilities as weights.

4. Families of learnable concepts

In the following we shall assume that we have selected the bounded square loss function (6).

In order to calculate the probability in expression (5), let us consider the random variables

$$Z_x = \begin{cases} m(x) & \text{if } [q(x) - p(x)]^2 > \delta^2 \\ & \qquad\qquad ; x \in \Omega \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

and let us call

$$s(x) = \Pr[Z_x = m(x)] = 1 - \Pr[-\delta \leq q(x) - p(x) \leq \delta]. \tag{21}$$

Note that

$$L_1(q) = \sum_{x \in \Omega} Z_x \,. \tag{22}$$

The mean values of Z_x and $L_1(q)$ are

$$E[Z_x] = m(x)s(x)\,; \quad E[L_1(q)] = \sum_{x \in \Omega} m(x)s(x). \tag{23}$$

With this, (5) becomes

$$\Pr[L_1(q) \leq \varepsilon] \geq 1 - \alpha. \tag{24}$$

Taking into account the Markov inequality

$$\Pr[L_1(q) \leq \varepsilon] \geq 1 - \frac{E[L_1(q)]}{\varepsilon} = 1 - \varepsilon^{-1} \sum_{x \in \Omega} m(x)s(x)$$

it follows that if we can prove for $n > N$ (or $\tau > T$)

$$\sum_{x \in \Omega} m(x)s(x) \leq \varepsilon\alpha \tag{25}$$

then our concept can be learnt in time N (or T).

The approximation of $s(x)$ is based on the following properties of the relative frequency. Let A be one of the possible outcomes of an experiment, and $1 > p = P(A) > 0$, and denote by $\zeta_n(p)$ the relative frequency of A after a sequence of n independent experiments. From Bernstein's improvement of the Chebyschev inequality we get

Theorem 1. *For* $0 < \gamma < p(1-p)$

$$\mathrm{P}[|\zeta_n(p) - p| \geq \gamma] < 2\exp\left[-\frac{n\gamma^2}{2p(1-p)\left(1 + \frac{\gamma}{3p(1-p)}\right)}\right].$$

From this theorem easily follows that for $\beta > 0$, $0 < \gamma < p(1-p)$ and

$$n > \frac{2\log\left(\frac{2}{\beta}\right)}{3\gamma^2} = n_0(\gamma, \beta)$$

we have

$$\mathrm{P}[|\zeta_n(p) - p| \geq \gamma] \leq \beta\,. \tag{26}$$

In order to calculate $s(x)$ we shall modify expression (19) in the following way:

$$q(x) = \begin{cases} \delta & \text{if} \quad \dfrac{n_x^+}{n_x} < \delta + \dfrac{\delta}{2}, \\ \dfrac{n_x^+}{n_x} & \text{if} \quad \delta + \dfrac{\delta}{2} \leq \dfrac{n_x^+}{n_x} \leq 1 - \delta - \dfrac{\delta}{2}, \\ 1 - \delta & \text{if} \quad \dfrac{n_x^+}{n_x} > 1 - \delta - \dfrac{\delta}{2} \end{cases} \tag{27}$$

if $n_x \geq n_0\left(\dfrac{\delta}{2}, \delta_1\right)$.

The estimation of $s(x)$ from above is based on the inequalitiy

$$\begin{aligned} s(x) &= \Pr[|q(x) - p(x)| > \delta] \leq \\ &\leq \Pr\left[|q(x) - p(x)| > \delta \mid n_x \geq n_0\left(\frac{\delta}{2}, \beta_1\right)\right] + \\ &+ \Pr\left[n_x < n_0\left(\frac{\delta}{2}, \beta_1\right)\right]. \end{aligned}$$

We shall prove that using (27) the following estimation

$$s(x) \leq \beta_1 + \Pr\left[n_x < n_0\left(\frac{\delta}{2}, \beta_1\right)\right] \tag{28}$$

holds for all the three cases in (27).

Case (a):

$$|q(x) - p(x)| = |\delta - p(x)| \quad \text{if} \quad \frac{n_x^+}{n_x} < \delta + \frac{\delta}{2}\,.$$

So, if $p(x) < 2\delta$ no contribution to $s(x)$ exists.
For $p(x) \geq 2\delta$ from inequality (26)

$$\begin{aligned} &\Pr\left[|q(x) - p(x)| > \delta \,\middle|\, n_x \geq n_0\left(\frac{\delta}{2}, \beta_1\right)\right] = \\ &= \Pr\left[p(x) - \frac{n_x^+}{n_x} > \frac{\delta}{2} \,\middle|\, n_x \geq n_0\left(\frac{\delta}{2}, \beta_1\right)\right] \leq \beta_1\,. \end{aligned}$$

Thus, (28) is satisfied for all $0 \leq p(x) \leq 1$.

Case (b):

$$|q(x) - p(x)| = \left|\frac{n_x^+}{n_x} - p(x)\right| \quad \text{if} \quad \delta + \frac{\delta}{2} \leq \frac{n_x^+}{n_x} \leq 1 - \delta - \frac{\delta}{2}.$$

From inequality (26), if $\delta \leq p(x) \leq 1 - \delta$

$$\Pr\left[|q(x) - p(x)| > \delta \,\middle|\, n_x \geq n_0\left(\frac{\delta}{2}, \beta_1\right)\right] \leq$$
$$\Pr\left[\left|\frac{n_x^+}{n_x} - p(x)\right| \geq \frac{\delta}{2} \,\middle|\, n_x \geq n_0\left(\frac{\delta}{2}, \beta_1\right)\right] \leq \beta_1.$$

If $p(x) < \delta$, then

$$\Pr\left[|q(x) - p(x)| > \delta \,\middle|\, n_x \geq n_0\left(\frac{\delta}{2}, \beta_1\right)\right] \leq$$
$$\Pr\left[\zeta_{n_x}\left(p(x)\right) > \delta + \frac{\delta}{2} \,\middle|\, n_x \geq n_0\left(\frac{\delta}{2}, \beta_1\right)\right] \leq$$
$$\Pr\left[\zeta_{n_x}(\delta) > \delta + \frac{\delta}{2} \,\middle|\, n_x \geq n_0\left(\frac{\delta}{2}, \beta_1\right)\right] \leq \beta_1,$$

and similarly, if $p \geq 1 - \delta$ we have the same inequality.

Thus, (28) is satisfied for all $0 \leq p(x) \leq 1$.

Case (c):

$$|q(x) - p(x)| = |1 - \delta - p(x)| \quad \text{if} \quad \frac{n_x^+}{n_x} > 1 - \delta - \frac{\delta}{2}.$$

Now, similarly to Case(a), we have that (28) is satisfied for all $0 \leq p(x) \leq 1$.

From (19) and (28) we get

$$(29) \qquad \sum_{x \in \Omega_0 \cup \Omega_1} m(x)s(x) \leq \beta_1 + \sum_{x \in \Omega_0 \cup \Omega_1} m(x) Pr\left[n_x < n_0\left(\frac{\delta}{2}, \beta_1\right)\right].$$

If we now sort the values $m(x)$ into increasing sequence

$$m(x_1) \leq m(x_2) \leq \ldots \leq m(x_{|\Omega_0|+|\Omega_1|}),$$

and let for L

$$\sum_{i=1}^{L-1} m(x_i) < \beta_2 \leq \sum_{i=1}^{L} m(x_i) \tag{30}$$

then

$$m(x_i) \geq \frac{\beta_2}{L} \geq \frac{\beta_2}{|\Omega_0| + |\Omega_1|} \quad \text{for} \quad i \geq L.$$

From (29) and (30) we get

$$\sum_{x \in \Omega_0 \cup \Omega_1} m(x)s(x) < \beta_1 + \beta_2 + \sum_{i \geq L} m(x_i) Pr\left[n_{x_i} < n_0\left(\frac{\delta}{2}\beta_1\right)\right] \tag{31}$$

4.1. Multinomial case

If we are in the multinomial case, let μ denote a negative binomial, $NB(k, p)$, random variable, that is with probability mass function

$$P[\mu = k + 1] = \binom{k + \ell - 1}{k - 1} p^k (1 - p)^\ell \quad \text{for} \quad \ell = 0, 1, \ldots$$

Then, with $k = n_0\left(\frac{\delta}{2}, \beta_1\right)$ and $p = \frac{\beta_2}{|\Omega_0| + |\Omega_1|}$, since $m(x_i) \geq p$, we have

$$\Pr\left[n_{n_i} < n_0\left(\frac{\delta}{2}, \beta_1\right)\right] \leq \mathrm{P}[\mu > n] \quad \text{for} \quad i \geq L.$$

Using the Chebyschev inequality and $E(\mu) = k/p$ and $D^2(\mu) = k(1 - p)/p^2$, we get

$$\mathrm{P}[\mu > n] \leq \beta_3 \quad \text{for} \quad n > \frac{k}{p} + \beta_3^{-\frac{1}{2}} \frac{\sqrt{k(1-p)}}{p}$$

for this, it follows that for

$$n > n_0\left(\frac{\delta}{2}, \beta_1\right)(|\Omega_0| + |\Omega_1|)\beta_2^{-1}\left[1 + \beta_3^{-\frac{1}{2}} n_0^{-\frac{1}{2}}\left(\frac{\delta}{2}, \beta_1\right)\right] \quad \text{and} \quad \ell \geq L$$

we have

$$\Pr\left[n_{x_i} < n_0\left(\frac{\delta}{2}, \beta_1\right)\right] \leq \beta_3. \tag{32}$$

Combining (31) and (32) we get

$$\sum_{x \in \Omega_0 \cup \Omega_1} m(x)s(x) \leq \beta_1 + \beta_2 + \sum_{i \geq L} m(x_i)\beta_3 \leq \beta_1 + \beta_2 + \beta_3. \tag{33}$$

Now, choosing arbitrarily the weights

$$\gamma_1 + \gamma_2 + \gamma_3 = 1$$

and

$$\beta_1 = \gamma_1 \varepsilon\alpha\,, \quad \beta_2 = \gamma_2 \varepsilon\alpha\,, \quad \beta_3 = \gamma_3 \varepsilon\alpha$$

we satisfy (25) with

$$N \geq n_0\left(\frac{\delta}{2}, \gamma_1\varepsilon\alpha\right)(|\Omega_0| + |\Omega_1|)(\gamma_2\varepsilon\alpha)^{-1}\left(1 - (\gamma_3\varepsilon\alpha)^{-\frac{1}{2}} n_0^{-\frac{1}{2}}\left(\frac{\delta}{2}, \gamma_1\varepsilon\alpha\right)\right) =$$

$$= \frac{8\log\left(\frac{2}{\gamma_1\varepsilon\alpha}\right)}{3\delta^2\gamma_2\varepsilon\alpha}(|\Omega_0| + |\Omega_1|)\left(1 + \frac{\sqrt{3}\delta}{\sqrt{8\gamma_3\varepsilon\alpha\log\left(\frac{2}{\gamma_1\varepsilon\alpha}\right)}}\right) = \tag{34}$$

$$= \frac{|\Omega_0| + |\Omega_1|}{\delta\gamma_2\varepsilon\alpha}\left(\frac{8\log\left(\frac{2}{\gamma_1\varepsilon\alpha}\right)}{3\delta} + \frac{\sqrt{8\log\left(\frac{2}{\gamma_1\varepsilon\alpha}\right)}}{\sqrt{3\gamma_3\varepsilon\alpha}}\right).$$

From (34) we get the following theorem.

Theorem 2. *A general class of concepts F given by the parameters $|\Omega_0|$ and $|\Omega_1|$ is learnable iff there exists a polynomial $Q(x)$ such that $|\Omega_0| + |\Omega_1| \leq Q(t)$ for the whole class.*

The only if part of the theorem is obvious for the kind of concepts where $p(x)$ is restricted to be outside of the interval $\left[\frac{\delta}{2}, 1 - \frac{\delta}{2}\right]$ and $m(x) = (|\Omega_0| + |\Omega_1|)^{-1}$ for $x \in \Omega_0 \cup \Omega_1$.

4.2. Poissonian case

If we are in the Poissonian sampling case, a Gamma variable $\mu = G(k, \lambda)$ with

$$k = n_0\left(\frac{\delta}{2}, \beta_1\right) \quad \text{and} \quad \lambda = \frac{\beta_2}{|\Omega_0| + |\Omega_1|}$$

must be used instead of a Negative Binomial random variable. Then from the Chebyschev inequality and $E[\mu] = k/\lambda$ and $D^2(\mu) = k/\lambda^2$ we get

$$\mathrm{P}[\mu < n] \leq \beta_3 \quad \text{for} \quad n > \frac{k}{\lambda} + \beta_3^{-\frac{1}{2}}\frac{\sqrt{k}}{\lambda}$$

which also leads to (34).

5. Examples of application

In order to illustrate the application of the above results, we give the following two examples.

Example 1. Assume a population of patients (objects) in a hospital which are defined in terms of three different binary symptoms (feature variables): "pain", "weight loss" and "vomits". We define the concept "gastric adenocarcinoma" and we find that the probabilities $m(x)$ and $p(x)$ are those given in Table 1, where P, W and V refer to pain, weight loss and vomits, respectively. Figure 1 shows the gastric adenocarcinoma concept (shaded region) and the values of $m(x)p(x)$ and $m(x)[1 - p(x)]$ associated with the corresponding combinations of symptoms.

$x = (P, W, V)$	$m(x)$	$p(x)$
(0,0,0)	0.06	0.001
(0,0,1)	0.14	0.002
(0,1,0)	0.06	0.002
(0,1,1)	0.14	0.998
(1,0,0)	0.09	0.002
(1,0,1)	0.21	0.998
(1,1,0)	0.09	0.997
(1,1,1)	0.21	0.999

Table 1. Values of $m(x)$ and $p(x)$ for the adenocarcinoma concept

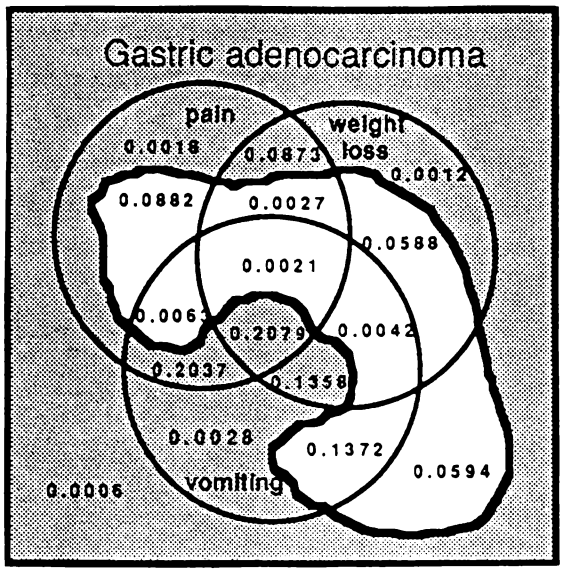

Figure 1. Adenocarcinoma concept and values of $m(x)p(x)$ *and* $m(x)[1-p(x)]$

If we want to learn a concept like gastric adenocarcinoma, the δ-error need not be related to the lowest or highest level of $p(x)$. It is related to the level of the low and high probabilities. If it is important to distinguish $p(x) = 0.01$ from $p(x) = 0.0001$, then we need $\delta < 0.005$. It depends, of course, on the medical doctor whether he considers $p(x) = 0.01$ a low probability or not.

In this case $|\Omega_0| = 0$ and $|\Omega_1| = 8$ and we assume $\alpha = 0.05$, $\varepsilon = 0.05$ and $\delta = 0.02$. Using now expression (34) we get $n = 573878666$.

Example 2. Figure 2.a shows the security mechanism of a room which is composed of two subsystems. The first, C, consists of a video–camera which transmits the image to a computer for analysis. After the analysis, the computer decides whether or not to activate a relay which closes an electric circuit with a battery activating an alarm. The second, F, consists of a photoelectric cell, D, which closes another electric circuit E with an alarm activated by a battery. Figure 2.b shows the rules associated with the alarm system. Note that the first system has been simplified to hardware plus software, and that rules are interpreted in a weak sense (conclusions are very likely but not sure).

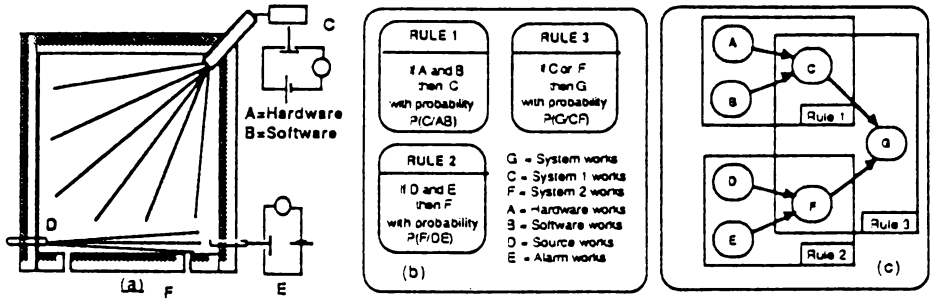

Figure 2. Security system: rules and influence diagram

Table 2 shows the probability distribution $m(x)$ where $x \in \{0,1\}^4$. The components of vector x are associated with components A, B, D and E. The value 1 indicates that the associated element works correctly and the value 0 that it fails to work. Probabilities $p_1(x)$, $p_2(x)$ and $p_3(x)$ define the concepts C, F and G, respectively.

$x = (A, B, D, E)$	$m(x)$	$p_1(x)$	$p_2(x)$	$p_3(x)$
(0,0,0,0)	0.0144	0.001	0.000	0.00
(0,0,0,1)	0.0216	0.001	0.002	0.00
(0,0,1,0)	0.0336	0.002	0.000	0.00
(0,0,1,1)	0.0504	0.001	0.998	1.00
(0,1,0,0)	0.0336	0.001	0.000	0.00
(0,1,0,1)	0,0504	0,002	0,000	0.00
(0,1,1,0)	0.0784	0.001	0.001	0.00
(0,1,1,1)	0.1176	0.001	1.000	1.00
(1,0,0,0)	0.0216	0.001	0.000	0.00
(1,0,0,1)	0,0324	0.001	0.000	0.00
(1,0,1,0)	0.0504	0.001	0.001	0.00
(1,0,1,1)	0.0756	0.001	1.000	1.00
(1,1,0,0)	0.0504	0.998	0.001	1.00
(1,1,0,1)	0.0756	0.998	0.000	1.00
(1,1,1,0)	0.1176	0.999	0.000	1.00
(1,1,1,1)	0.1764	0.998	0.997	1.00

Table 2. Values of $m(x)$ and values of $p(x)$ for three different concepts

The required samples sizes to learn the concepts C, F and G for $\alpha = 0.05$, $\beta = 0.05$ and $\delta = 0.02$ are $n = 11477573329$. Note that here we have $\{|\Omega_0| =$

$0, |\Omega_1| = 16\}$, $\{|\Omega_0| = 10, |\Omega_1| = 6\}$ and $\{|\Omega_0| = 16, |\Omega_1| = 0\}$, for concepts C, F and G respectively.

This security mechanism is the implementation of the Boolean function $(x_1 \wedge x_2) \vee (x_3 \wedge x_4)$ with some small probabilistic bias Δ. This means $p(x) < \Delta$ corresponds to a false value, $1 - p(x) < \Delta$ corresponds to a true value. Now, learning means: find a Boolean function acceptable with Δ bias, or show some significant error. The acceptable bias Δ is limited by the δ parameter of the learning algorithm. For example choosing a small bias by $\delta = 0.0001$ (which is not small for testing an integrated circuit), the volume of the required sample size becomes extremely large: $n \geq 39880407899885$.

From our numerical examples it turns out that the required sample size might be extremely large for practical application. So we have to give a warning for applications: in spite of the polynomiality the accuracy and the level of the learning cannot be chosen arbitrarily. When the number of observations is limited, we cannot continue the automatic, algorithmic learning, the experts have to search for new feature variables.

References

[1] **Bishop Y.M.M., Fienberg S.E. and Holland P.W.**, *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, Mass., The MIT Press, 1975.

[2] **Pitt L. and Valiant L.G.**, Computational Limitations on Learning from Examples. *Journal of the Association for Computing Machinery*, **35** (4) (1988), 965-984.

[3] **Valiant L.G.**, A Theory of the Learnable. *Communications of the ACM*, **27** (11) (1984), 1134–1142.

(Received April 30, 1991)

E. Alvarez, E. Castillo and J.M. Sarabia
Department of Applied Mathematics and Computational Sciences
University of Cantabria
39005 Santander, Spain

A. Benczúr
Computer Center
Eötvös Loránd University
H-1117 Budapest, XI. Bogdánfy u. 10/b.
Hungary