CONVERGENCE PROPERTIES OF A LEARNING ALGORITHM[1]

By Leo Breiman and Zivia S. Wurtele

University of California, Los Angeles

1. Introduction. In a recent paper Albert [1] introduced an algorithm for learning to classify individuals that are drawn from a population which is partitioned into two categories. The purpose of this note is to discuss an algorithm which is simpler, in the sense that at any given stage half as many items are retained in memory.

In the learning process described by the algorithm, observations are made on individuals one at a time and the current estimate of the required partitioning may be adjusted after each observation, on the basis of knowledge of the category to which the individual observed belongs. At any given time, the current estimate of the partitioning is all that is held in memory; past history is lost except insofar as it has been incorporated into the present estimate. The learning process of perceptrons, as well as that of other artificial intelligences, is of this general form.

2. Notation and assumptions. It is assumed that each individual is a member of one and only one of two categories. The results obtained are applicable to the more general case, however, for they may be applied to appropriate partitions of a set of three or more categories into two subsets.

Each individual in the population is characterized by an attribute vector X in m-dimensional Euclidean space; let S_1, S_2 be the sets of vectors attributed to members of the first and second categories, respectively. We shall suppose that this characterization is *sufficiently rich* with respect to the given classification problem, that is to say that the regions S_1 and S_2 are separable by a hyperplane (except for a set of probability measure zero). This terminology is appropriate to situations for which in the case of failure of the condition of sufficient richness, a re-examination of the world of individuals and the subsequent increasing of the number of components of the characterizing vectors can be expected to yield a new description for which this condition is satisfied. The question of whether, in a particular case, a sufficiently rich characterization can be achieved is obviously crucial but beyond the scope of this paper.

It is supposed that initially there are two samples: $X_1^{(1)}, \cdots, X_p^{(1)}$ from S_1 and $X_1^{(2)}, \cdots, X_q^{(2)}$ from S_2. Let X_n be the nth vector sampled after the initial $p + q$ vectors. We assume the following.

Assumption. There exist two bounded sets S_1, S_2 and a distribution Q on $S_1 \cup S_2$ such that

Received 31 October 1963; revised 8 May 1964.

[1] This work was supported partly by the Office of Naval Research under Task 047-041, and partly by the Western Management Science Institute under a grant from the Ford Foundation. Reproduction in whole or in part is permitted for any purpose of the United States Government.

(a) *There is a vector* B^* *and a* $\delta > 0$ *such that* $Q(x; (B^*x) > \delta, x \varepsilon S_1) = 1$ *and* $Q(x; (B^*x) < -\delta, x \varepsilon S_2) = 1$.

(b) *Samples* $X_1, X_2, \cdots$ *are drawn independently from the distribution* Q.

(c) *For* $i = 1, 2, Q(S_i) > 0$.

3. The algorithm. Estimate B^* as follows.

(1) Let the initial estimate of B^* be $B_1 = \sum_{i=1}^{p} \theta_i^{(1)} X_i^{(1)} - \sum_{i=1}^{q} \theta_i^{(2)} X_i^{(2)}$, where the $\theta_i^{(j)}$'s are non-negative but not all zero.

(2) $B_{n+1} = B_n + e_n^{(1)} X_n - e_n^{(2)} X_n$, where the e's are determined in accordance with the following rules.

(a) If $X_n \varepsilon S_1$ then

$$e_n^{(2)} = 0, e_n^{(1)} = 0 \text{ if } (B_n X_n) \geqq 0, \quad e_n^{(1)} = -(B_n X_n)/|X_n|^2 \text{ if } (B_n X_n) < 0.$$

(b) If $X_n \varepsilon S_2$, then

$$e_n^{(1)} = 0, e_n^{(2)} = 0 \text{ if } (B_n X_n) \leqq 0, e_n^{(2)} = (B_n X_n)/|X_n|^2 \text{ if } (B_n X_n) > 0.$$

This algorithm may be described geometrically as follows. If X_n is not oriented correctly with respect to the plane $(B_n X) = 0$, then $B_{n+1} = B_n + U_n$, where U_n is the shortest vector which can be added to B_n so that $(B_{n+1} X_n) = 0$; otherwise, $B_{n+1} = B_n$.

The results below also hold for the class of algorithms defined as follows: $e_i^{(j)} = 0$ in accordance with the algorithm above; if $e_i^{(j)} \neq 0$, $\theta |(B_n X_n)|/|X_n|^2 \leqq e_i^{(j)} \leqq 2|(B_n X_n)|/|X_n|^2$, where θ is a fixed number in the interval $(0, 1)$.

4. Convergence properties of the algorithm. A convergence theorem analogous to the one below was proved by Papert [2] for the case where the attribute vectors X are binary. Note that

$$B_{n+1} = B_1 + \sum_{i=1}^{n} e_i^{(1)} X_i - \sum_{i=1}^{n} e_i^{(2)} X_i .$$

Assume none of the X_i fall in the exceptional set of probability zero violating assumption (a). We introduce coordinates as follows. Let E be a unit vector in the B^* direction. Write $X_n = Y_n + \kappa_n E$, where Y_n is orthogonal to E, and $B_n = C_n + \beta_n E$, where C_n is orthogonal to E. Let $\epsilon = \delta/|B^*|$. If $X_n \varepsilon S_1$, then $\kappa_n > \epsilon$; and if $X_n \varepsilon S_2$, $\kappa_n < -\epsilon$.

LEMMA 1. $\beta_{n+1} \geqq \beta_n \geqq \beta_1 > 0$.

PROOF. If $X_n \varepsilon S_1$, $\beta_{n+1} = \beta_n + e_n^{(1)} \kappa_n$ and $e_n^{(1)} \kappa_n \geqq 0$. If $X_n \varepsilon S_2$, $\beta_{n+1} = \beta_n - e_n^{(2)} \kappa_n$, and $e_n^{(2)} \kappa_n \leqq 0$. In either case $\beta_{n+1} \geqq \beta_n$. Similarly, since $B_1 = \sum_{i=1}^{p} \theta_i^{(1)} X_i^{(1)} - \sum_{i=1}^{q} \theta_i^{(2)} X_i^{(2)}$, where the $\theta_i^{(j)}$'s are non-negative, β_1 is positive.

LEMMA 2. $|B_{n+1}| \leqq |B_n|$.

PROOF. If $e_n^{(1)} \neq 0$, $|B_{n+1}|^2 - |B_n|^2 = (e_n^{(1)})^2 |X_n|^2 + 2e_n^{(1)}(B_n X_n)$. This is negative since $e_n^{(1)} \leqq -2(B_n X_n)/|X_n|^2$.

If $e_n^{(2)} \neq 0$, $|B_{n+1}|^2 - |B_n|^2 = (e_n^{(2)})^2 |X_n|^2 - 2e_n^{(2)}(B_n X_n)$. This is negative, since $e_n^{(2)} \leqq 2(B_n X_n)/|X_n|^2$.

LEMMA 3. $B_n \to \tilde{B}$, *a finite, non-zero vector.*

PROOF. Note that $|B_n|^2 = \beta_n^2 + |C_n|^2$. Since by the previous lemmas, $|B_n|$ is non-increasing and β_n is non-decreasing, it follows that $\beta_n \to \tilde{\beta} < \infty$. (Observe also that $|C_n|$ must be non-increasing.) But

$$\beta_{n+1} = \beta_1 + \sum_{i=1}^{n} e_i^{(1)}\kappa_i - \sum_{i=1}^{n} e_i^{(2)}\kappa_i \geqq \beta_1 + \epsilon \sum_{i=1}^{n} e_i^{(1)} + \epsilon \sum_{i=1}^{n} e_i^{(2)}.$$

Thus, since $\epsilon > 0$, $\sum e_i^{(1)}$ and $\sum e_i^{(2)}$ must converge to finite limits. Since S_1 and S_2 are bounded, we have

$$B_{n+1} = B_1 + \sum_{i=1}^{n} e_i^{(1)}X_i - \sum_{i=1}^{n} e_i^{(2)}X_i \to \tilde{B},$$

a finite non-zero vector.

THEOREM. *With probability one,* (a) $Q(x; (\tilde{B}x) < 0, x \varepsilon S_1) = 0$ *and* (b) $Q(x; (\tilde{B}x) > 0, x \varepsilon S_2) = 0$.

PROOF. We first prove (a). From Lemma 3 we know that $\sum_{n=1}^{\infty} e_n^{(1)}$ is finite and therefore that $e_n^{(1)} \to 0$. Thus for every $\theta > 0$, $\lim_{n\to\infty} P(e_n^{(1)} > \theta) = 0$. For every $\alpha > 0$, since the event $(B_nX_n) < -\alpha/2$ is contained in the event $e_n^{(1)} > \alpha/2h^2$, where h is an upper bound of $|X|$ on $S_1 \cup S_2$, it follows that

$$\lim_{n\to\infty} P((B_nX_n) < -\alpha/2, X_n \varepsilon S_1) = 0.$$

Since the X's are independent,

$$P((B_nX_n < -\alpha/2, X_n \varepsilon S_1 \mid X_1, \cdots, X_{n-1}) = Q(x; (B_nx) < -\alpha/2, x \varepsilon S_1).$$

Thus $P((B_nX_n) < -\alpha/2, X_n \varepsilon S_1) = \int Q(x; (B_nx) < -\alpha/2, x \varepsilon S_1)\, dP$. Let $M_n = \min [\inf_{x\varepsilon S_1}((\tilde{B} - B_n)x), 0]$, and observe that $M_n \leqq 0$ and that $M_n \to 0$ almost surely. Now

$$Q(x; (\tilde{B}x) < -\alpha/2 + M_n, x \varepsilon S_1) \leqq Q(x; (B_nx) < -\alpha/2, x \varepsilon S_1).$$

Therefore,

$$\int Q(x; (\tilde{B}x) < -\alpha/2 + M_n, x \varepsilon S_1)\, dP \to 0.$$

Note that

$$\begin{aligned}\int Q(x; (\tilde{B}x) &\leqq -\alpha/2 + M_n, x \varepsilon S_1)\, dP \\ &\geqq \int_{M_n < -\alpha/2} Q(x; (\tilde{B}x) < -\alpha/2 + M_n, x \varepsilon S_1)\, dP \\ &\quad + \int_{M_n \geqq -\alpha/2} Q(x; (\tilde{B}x) < -\alpha, x \varepsilon S_1)\, dP \\ &\geqq -P(M_n < -\alpha/2) + \int Q(x; (\tilde{B}x) < -\alpha, x \varepsilon S_1)\, dP.\end{aligned}$$

Since $P(M_n < -\alpha/2) \to 0$, we have $Q(x; (\tilde{B}x) < -\alpha, x \varepsilon S_1) = 0$ almost surely for any $\alpha > 0$. Part (a) follows from the fact that $Q(x; (\tilde{B}x) < \gamma)$ is continuous from the left in γ. Similarly, for part (b).

5. Conclusions. It has been assumed that sampling is random from the entire population of individuals to be classified. If stratified sampling, i.e., by categories, is permitted, convergence may be made more rapid. It may be feasible, for example, to alternate categories by sampling from a given category as long as the vector sampled requires an adjustment in the estimate of the dividing hyperplane, and as soon as a vector is obtained which is correctly oriented with respect to the hyperplane, switching to the alternative category. Furthermore, if any information about the distribution of X is available, it might be practicable to incorporate it into the stratified sampling plan.

6. Acknowledgments. We are grateful to Professors T. Ferguson and J. MacQueen for discussions of this paper.

REFERENCES

[1] ALBERT, A. (1963). A mathematical theory of pattern recognition. *Ann. Math. Statist.* **34** 284–299.

[2] PAPERT, S. (1961). Some mathematical models of learning. *Information Theory, The Fourth London Symposium*, (Edited by Colin Cherry). Butterworth, London and Washington.