CAPACITY PROBLEMS FOR LINEAR MACHINES*

Thomas M. Cover

Stanford University, Stanford, California

THIS PAPER is concerned with some problems of a combinatorial geometric nature which are related to the general problem of describing the behavior and capacity of networks of linear threshold devices.

We shall first review the problems of counting the number of linearly separable dichotomies and counting the number of linearly inducible orderings of n points in d-space. We shall use the form of the solutions of these two problems to guess the solution to a third—that of counting the number of linearly inducible r-chotomies of n points in d-space. During this development we shall mention several straightforward techniques for solving linear orderings.

Finally, we shall investigate the computational capacity, as a function of the number of variable weights, of arbitrary networks of linear threshold devices. All of our results tend to indicate that the pattern-classifying capacity of networks of linear threshold devices is of the same order of magnitude as the number of variable weights. Hence, some idea is given of the number of patterns necessary to train such systems.

The Linear Dichotomization Problem

Let $C(n,d)$ be the number of ways in which n patterns in general position in d-space may be partitioned into two sets X_1 and X_2 by the assignment function

*This work has been partially supported under contract AF 49 (638) 1517.

$$x \epsilon X_1, \text{ if } w^t x \geq 0$$
$$x \epsilon X_2, \text{ if } w^t x < 0$$

where w is a freely chosen weight vector in Euclidean d-space.

This problem has a long history of contributions,[1-5] culminating in Winder's[5] very nice generalization to points not in general position.

It can be shown that $C(n,d)$ satisfies the recurrence relation

$$C(n,d) = C(n\text{-}1,d) + C(n\text{-}1,d\text{-}1)$$

with the boundary conditions

$$C(n,d) = 2, n = 1, 2, \ldots$$
$$C(1,d) = 2, d = 1, 2, \ldots$$

yielding the solution

$$C(n,d) = 2\sum_{k=0}^{d-1} \binom{n-1}{k}$$

where $$\binom{n-1}{k} = (n-1)!/k!\,(n-1-\mathrm{k})!$$

Thus $C(n,d)$ is independent of the precise configuration of the patterns up to general position.

The Linear Ranking Problem

Consider a collection of n pattern vectors $x_1, x_2, \ldots, x_n$ in E^d which are ranked according to their orthogonal projection onto a reference vector $w \epsilon E^d$. If π is a permutation of the integers $1,2,\ldots,n$ we shall say that the ranking π is linearly inducible if there exists a weighting vector $w \epsilon E^d$ such that

$$w^t x_{\pi(1)} > w^t x_{\pi(2)} > \ldots > w^t x_{\pi(n)}$$

Let $Q(n,d)$ be the number of linearly inducible rankings of n pattern vectors in E^d. If the patterns are in general position it has been shown[6,7,8] that $Q(n,d)$ satisfies the recurrence relation

$$Q(n,d) = Q(n\text{-}1, d) + (n\text{-}1)\, Q\,(n\text{-}1,d\text{-}1)$$

with the boundary conditions

$$Q(n,1) = 2 \qquad n = 2,3,\ldots$$
$$Q(2,d) = 2 \qquad d = 1,2,\ldots$$

This relation was first surmised by Bennet[6] and was elaborated upon in Bennet and Hayes[7] in connection with the problem of determining the minimum natural dimension for a set of data points for which an ordering of the interpoint distances is specified. Their proof, which is heuristic in nature, is completed and established rigorously in Cover[8], and an explicit specification of general position (with respect to the linear ordering problem) is provided as a consequence.

This system has the solution

$$Q(n,d) = 2\sum_{k=0}^{d-1}\binom{n-1}{k}^* \qquad \begin{matrix} d=1,2,\ldots \\ n=2,3,\ldots \end{matrix}$$

where

$$\binom{n-1}{k}^* = \sum_{i_1, i_2 \ldots i_k}$$

where the summation is over

$$1 < i_1 < i_2 < \ldots < i_k < n$$

Here, as in the case of $C(n,d)$, we see that $Q(n,d)$ is independent of the configuration of the pattern vectors up to general position.

Achieving Linear Rankings

Before proceeding to the r-chotomization problem in the next section, we should like to mention some simple methods for the determination of a weight vector w^* achieving a desired linear ordering π.

Consider the algorithm which, upon being presented a pair of pattern x_i and x_j at the kth stage, increments the weight vector w_k by $x_i - x_j$ only if w_k incorrectly orders x_i and x_j. Specifically, for $\pi(i) < \pi(j)$,

$$w_{k+1} = \begin{cases} w_k + x_i - x_j, & w_k^t x_i < w_k^t x_j \\ w_k, & w_k^t x_i > w_k^t x_j \end{cases}$$

Then, by the perceptron convergence algorithm, for any sequence of pairs of vectors from $x_1, x_2, \ldots, x_n$ the sequence of weight vectors w_k will make only a finite number of mistakes. Moreover, if a sequence of pairs of vectors is presented in which each pair occurs infinitely often (for example, if we run serially through the set of all pairs), then $\{w_k\}$ converges, in a finite number of corrections, to a vector w^* which yields the desired ordering.

If an orderly procedure like the fixed increment, relaxation, or simplex method is to be used for training, an important saving in time will be effected if only the $n-1$ *extreme* pattern pairs $(x_{i_1} - x_{i_2})$, $(x_{i_2} - x_{i_3}), \ldots,$ $(x_{i_{n-1}} - x_{i_n})$ corresponding to the ordering $(i_1, i_2, \ldots, i_n)$ are used for training. Since all other inequalities are consequences of these, the number of patterns in the training set is reduced from $\binom{n}{2}$ to $n-1$.

THE r-CATEGORY LINEAR ASSIGNMENT PROBLEM

Consider using hyperplanes to partition a pattern set $\{x_1, x_2, \ldots, x_n\}$ into r categories.

Let $X_1, X_2, \ldots, X_r$ be a partition of $\{x_1, x_2, \ldots, x_n\}$. Then several definitions of r-separability seem to be natural in the context of this discussion. For example, we may say that $X_1, X_2, \ldots, X_r$ is r-*separable* if there exist r-1 hyperplanes which partition d-space into cells such that no two patterns from different X_i's lie in the same cell. Or we may restrict the hyperplanes to be parallel in the above definition. We may even wish to add the constraint that each of the X_i's be completely contained in a single cell of the partition formed by the hyperplanes. Clearly there are other possible natural definitions. Unfortunately, we do not know at this time which definition leads to the number of r-chotomies $C_r(n,d)$ which we shall now develop.

Consider n points in general position in d-space. Since the number of linearly inducible dichotomies $C(n,d)$ satisfies

$$C(n,d) = C(n\text{-}1,d) + C(n\text{-}1, d\text{-}1)$$

and since the number of linearly inducible orderings $Q(n,d)$ satisfies

$$Q(n,d) = Q(n\text{-}1,d) + (n\text{-}1)\, Q(n\text{-}1,d\text{-}1)$$

it seems, by analogy, that the number of linearly inducible r-chotomies $C_r(n,d)$ should satisfy

$$C_r(n,d) = C_r(n\text{-}1,d) + (r\text{-}1)\, C_r(n\text{-}1,d\text{-}1)$$

Examining this relation, we find, assuming natural boundary conditions, that

$$C_r(n,d) = r \sum_{k=0}^{d-1} \binom{n\text{-}1}{k} (r\text{-}1)^k$$

Since there are r^n possible partitions of n points into r sets, we see that the probability that a "random" r-chotomy is linearly inducible is

$$C_r(n,d)/r^n = \sum_{k=0}^{d-1} \binom{n\text{-}1}{k} (1\text{-}1/r)^k\, (1/r)^{n-k-1}$$

which is just the probability that d-1 or fewer successes result from n-1 tosses of a coin with bias (1-1/r). Since the expected number of such success is $n(1\text{-}1/r)$, it is clear that $C_r(nd)/r^n$ is near 1 or 0 accordingly as $d > n(1\text{-}1/r)$ or $d < n(1\text{-}1/r)$.

Thus, it would be natural to define $n^* = dr/(r-1)$ to be the pattern classifying capacity of such a system. This value for the capacity agrees precisely with that conjectured by Brown[9] and is supported by empirical evidence gathered by Brown. In the special case $r = 2$, it agrees with the definition of capacity in Ref. 3. However, the particular linear threshold system yielding $C_r(n,d)$ remains unknown.

Minimum Complexity of a Network

In this section, the material on the linear dichotomization problem will be applied to a large class of networks of linear threshold units in order to place a lower bound on the number of variable weights in a universal network. A network will be called *universal* with respect to a set of N pattern vectors if the network can implement each of the 2^N functions from the pattern set to $\{-1,1\}$. Cameron,[10] Winder,[11] and Joseph,[12] have studied several specific network organizations of linear threshold units and have determined lower bounds on the number of *linear threshold units* (gates) in a universal network.

It is known[3] that a single linear threshold unit has a capacity of two patterns per variable weight. Hence it is natural to ask for the capacity of a network of linear threshold units in terms of the total number of *variable weights.*

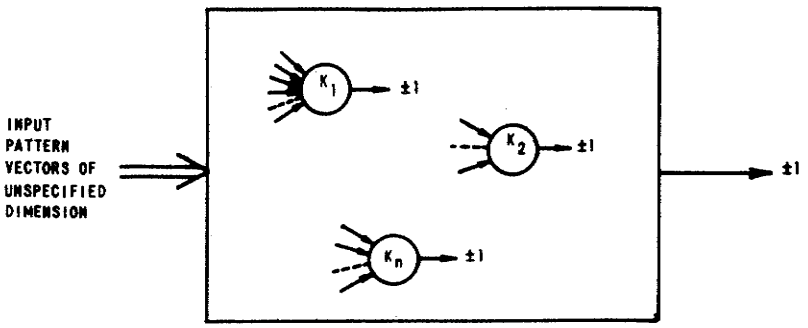

Fig. 1. Network of linear threshold units imbedded in arbitrary but fixed circuitry.

Consider a class **A** of networks on linear threshold units imbedded in fixed but arbitrary circuitry, as depicted in Fig. 1. However, in order to avoid problems of timing and stability it is required that there be no feedback from the output of any linear threshold unit to its input. Order the linear threshold units in any way that is not inconsistent with the flow of signal from input to output of the network, and let K_i denote the number of variable weights in the i^{th} linear threshold unit. Note that it is not required that all the inputs to the i^{th} linear threshold unit be utilized, nor is it required that all the dimensions of the input patterns be accommodated by the network. Let T denote the total number of weights in a given network in the class **A**. That is,

$$T=\sum_i K_i$$

Proposition. If a network in **A** containing a total of T variable weights is universal with respect to an input set of N patterns, then

$$\mathrm{T} > \frac{\mathrm{N}}{1+\log_2 N}$$

Proof. An upper bound is to be placed on the number of states of any T-weight network in **A**. Consider the rth linear threshold unit, with the weight vectors of the first r-1 linear threshold units fixed. The rth unit receives input K_r-tuples $\{x_1', x_2', \ldots, x_N'\}$ corresponding to the set of N inputs $\{x_1, x_2, \ldots, x_N\}$ to the network. Then, from the section on the linear dichotomization problem there are at most $C(N,K_r)$ different states of the rth unit with respect to $\{x_1', x_2', \ldots, x_N'\}$. (There are precisely $C(N,K_r)$ states if every K_r-element subset of $\{x_1', x_2', \ldots, x_N'\}$ is linearly independent.) Hence, an upper bound $r(N,T)$ on the number of states of the network is

$$r(N,T)=\max \Pi\,(2N^{K_i})$$
$$\Sigma K_i = T$$

Consider the crude bound on $C(N,K)$ holding for all positive integers N and K:

$$C(N,K) = 2\sum_{m=0}^{K-1}\binom{N\text{-}1}{m} \leqslant 2\sum_{m=0}^{K-1} N^m \leqslant 2N^K$$

Thus

$$r(N,T) \leqslant (2N)^T$$

There are 2^N functions mapping $\{x_1, x_2, \ldots, x_N\}$ to $\{-1,1\}$. Thus the number of states of a universal network must exceed 2^N. That is,

$$(2N)^T > 2^N$$

or

$$T > \frac{N}{1+\log_2 N}$$

REFERENCES

1. J. Steiner, *Jour-Reine, Angew. Math. (Crelle)* (1926), pp. 349-64.
2. L. Schlafli, *Gessammelte Mathematische Abhandlungen*, Basel, Verlag Birkhauser, 1950, Vol. I, pp 209-12.
3. T. M. Cover, "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition," *IEEE Trans. on Electronic Computers*, EC-14 (1965), pp. 326-34.
4. R. O. Winder, "Threshold Logic," Ph.D. dissertation, Princeton University, 1962.
5. ———, "Partition of *n*-Space by Hyperplanes," *Siam J. Appl. Math.*, July 1965.
6. Joseph F. Bennet, "Determination of the number of independent parameters of a score matrix from the examination of rank orders," *Psychometrika*, Vol. 21, No. 4, December, 1956, pp. 383-393.
7. Joseph F. Bennet, and William L. Hays, "Multidimensional unfolding: determining the dimensionality of ranked preference data", *Psychometrika*, Vol. 25, No. 1, March 1960, pp. 27-43.
8. T. M. Cover, "The Number of Linearly Inducible Orderings of Points in *d*-Space," *Siam J. Appl. Math.* Vol. 15, No. 2, (March 1967) pp. 434-39.
9. R. J. Brown, "Adaptive Multiple-Output Threshold Systems and Their Storage Capacities," TR 677-1, Stanford Electronics Labs, Stanford University, June, 1964.
10. S. H. Cameron, Tech. Report 60-600, *Proceedings of the Bionics Symposium*, Wright Air Development Division, Dayton, Ohio, 1960, pp. 197-212.
11. R. O. Winder, "Bounds on Threshold Gate Realizability," *IEEE Trans. on Electronic Computers* (correspondence), EC-12 (Oct. 1963) pp. 561-64.
12. R. D. Joseph, "The Number of Orthants in *n*-Space Intersected by an *s*-Dimensional Subspace," Tech. Memo 8, Project PARA, Cornell Aeronautical Laboratory, Buffalo,