Monotonicity of Linear Separability Under Translation

ALFRED M. BRUCKSTEIN AND THOMAS M. COVER

Abstract—A set of \(n\) pattern vectors are given in \(d\)-space and classified arbitrarily into two sets. The sets of patterns are said to be linearly separable if there exists a hyperplane that separates them. We ask whether translation of one of these sets in an arbitrary direction helps separability. Sometimes yes and sometimes no, but yes on the average. The average is taken over all classifications of the patterns into two sets. In fact, we prove that the probability of separability increases as the translation increases. Thus, we conclude that if points are drawn equiprobably from densities \(f_0(x)\) and \(f_1(x) = f_0(x + tw)\) then the probability of linear separability is minimum at \(t = 0\) and increases with \(t\) for \(t > 0\).

Index Terms—Convex sets, linear separability, monotonicity, pattern classification.

I. INTRODUCTION

Consider the standard statistical pattern classification problem in which the classifications \(\theta_1, \theta_2, \ldots, \theta_n\) are independent identically distributed random variables with \(P[\theta_i = 0] = P[\theta_i = 1] = 1/2\), and the corresponding vector-valued observations \(x_1, x_2, \ldots, x_n \in R^d\) are conditionally independently drawn according to \(f_{\theta_i}(x)\), where \(f_0(x)\) and \(f_1(x)\) are known probability density functions. Thus, the probability density of the classified set \(\{x_i, \theta_i \}_{i=1}^{n}\) is \(2^{-n} \prod_{i=1}^{n} f_{\theta_i}(x_i)\). The realization \(\{(x_i, \theta_i)\}_{i=1}^{n}\) is called linearly separable if there exists a vector \(u\) and a constant \(T\) such that

\[
\begin{align*}
    u^T x_i &> T \quad \text{for } \theta_i = 1 \\
    u^T x_i &< T \quad \text{for } \theta_i = 0.
\end{align*}
\]

The following result is well known (see, e.g., [11], [12]).

Theorem 1: If \(f_0(x) = f_1(x)\) then

\[
\Pr \{(X_1, \theta_1) \}_{i=1}^{n} \text{ is linearly separable} = \frac{1}{2^{n-1}} \sum_{i=0}^{d} \binom{n-1}{i}.
\]

Note that the probability does not depend on the underlying density. The proof of this theorem is based on a purely geometric argument which provides the number of dichotomies that can be induced by hyperplanes on a set of points in \(R^d\) in general position. This also makes it clear why the result is distribution free.

We now consider densities that differ by a translation. Given \(w\), an arbitrary unit vector in \(R^d\), we prove the following.

Theorem 2: If \(f_1(x) = f_0(x + tw)\) then

\[
\Pr \{(X_1, \theta_1) \}_{i=1}^{n} \text{ is linearly separable} \geq \frac{1}{2^{n-1}} \sum_{i=0}^{d} \binom{n-1}{i}.
\]

Manuscript received April 9, 1984; revised January 15, 1985. Recommended for acceptance by Josef V. Kitter. This work was supported in part by the National Science Foundation under Grants ECS78-10003 and NSF-82-11568, and in part by the Joint Services Electronics Program under Contract DAAG29-81-K-0057.

A. M. Bruckstein is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305.

T. M. Cover is with the Department of Electrical Engineering and Statistics, Stanford University, Stanford, CA 94305.

0162-8828/85/0500-0355$01.00 © 1985 IEEE
is a monotonic nondecreasing function of $t > 0$.

II. PROOF OF THEOREM 2

We shall first take the probability out of the problem by a simple observation and prove a counterpart of Theorem 2 for a strictly geometrical setup. The probabilistic result will easily follow.

The process of generating the classified samples for translated underlying distributions is equivalent to choosing all the points according to a density $f_0(x)$, classifying them at random (i.e., by flipping a fair coin) and then translating the points of class 1 by the translation vector $tw$. Consider thus a set of $n$ points in $\mathbb{R}^d$ in general position, and all its $2^n$ subsets, $S_1, S_2, \ldots, S_{2^n} \subseteq \{x_1, x_2, \ldots, x_n\}$. Each subset defines a classification or dichotomy $(S_k, S^c_k)$ of the original set of points, where $S^c_k$ denotes the complement of $S$ with respect to $\{x_1, x_2, \ldots, x_n\}$. The set pair $(S_k, S^c_k)$ is said to be linearly separable if there exists a separating hyperplane, i.e., if there exists a vector $v$ and a constant $T$ such that

$$v^T x \geq T \quad \text{for} \quad x \in S,$$

$$v^T x < T \quad \text{for} \quad x \in S^c.$$

(4)

We now evaluate the number of linearly separable dichotomies of points which result when translating the subsets $S_k$ by $tw$ to form the set $S_k + tw = \{x_i + tw \mid x_i \in S_k\}$. Associate to each $S_k$ a translation separability indicator function, as follows

$$I_k(t) = \begin{cases} 1, & \text{if } (S_k + tw, S^c_k) \text{ is linearly separable} \\ 0, & \text{if } (S_k + tw, S^c_k) \text{ is not linearly separable.} \end{cases}$$

(5)

**Definition**: Let $\Pi(t)$ be the number of linearly separable sets of points among $(S_k + tw, S^c_k)$, for $k = 1, 2, \ldots, 2^n$. It follows from (5) that

$$\Pi(t) = \sum_{k=0}^{2^n} I_k(t).$$

(6)

Suppose that $(S_k, S^c_k)$ is not linearly separable. Then translation of the points of $S_k$ can only make the dichotomy separable. If, however, we start with a separable dichotomy of points, translation may at some "critical" distance $t_c$ produce a nonseparable dichotomy. This happens when the convex hull of the points in $S_k + tw$ intersects the convex hull of the "static" points $S^c_k$. As $t \to \infty$, the dichotomy will again become separable (see Fig. 1). In Fig. 1(a), we see that $S_k = \{x_1, x_4, x_5, x_6, x_9\}$ becomes separable from $S^c_k$ at translation $t = t_1$, and remains separable thereafter. Fig. 1(b) shows $S_1 = \{x_1, x_4, x_5, x_7\}$ becoming nonseparable at $t = t_c$ and regaining separability for $t > t_c$. Thus it is clear that the count function $\Pi(t)$ is piecewise constant. Also, if $\{x_1, x_2, \ldots, x_n\}$ is in general position, then

$$\Pi(0) = 2^d \sum_{i=0}^{d-1} \binom{n-1}{i-1} \quad \text{and} \quad \Pi(\infty) = 2^n.$$

(7)

We shall argue that the following result holds,

**Geometric Theorem**: If $\{x_1, x_2, \ldots, x_n\}$ is in general position, the right-continuous version of $\Pi(t)$ is a piecewise constant nondecreasing function of $t$, for $t > 0$.

**Proof**: It is enough to consider the behavior of $\Pi(t)$ for $n > d + 1$ because $\Pi(t) = \Pi(0) = 2^n$ for all $t$ when $n \leq d + 1$. Showing that $\Pi(t)$ is nondecreasing is equivalent to proving that no "down-jumps" will occur. Since $\Pi(t) = \Sigma I_k(t)$ and the $I_k(t)$ are not necessarily monotonic, we wish to find for every $I_{p}(t)$ having a down-jump at $t_c$, another $I_{q}(t)$ having a cancelling up-jump at $t_c$. Suppose that $(S_p + tw, S^c_p)$ becomes nonseparable at $t = t_c$. This happens because (at least) one of the points of one subset crosses a face of the convex hull of the other [Fig. 2(a)]. Thus at the critical translation we will have a "collision hyperplane" on which there are points of both $S_p + tw$ and $S^c_p$. Now let $S_q$ be the uniquely defined subset comprising points which either belonged to $S_p$ and when translated by $tw$ moved to the collision plane or belonged to $S^c_p$ and were not on the collision plane. It is easy to see that at the critical translation $t_c$, $(S_q + tw, S^c_q)$ will change from nonseparable to separable and thus $I_p(t) + I_q(t)$ will not have a down-jump at $t_c$ [Fig. 2(a)]. Thus we have identified a set $S_q$ such that when $I_p(t)$ has a down-jump, $I_q(t)$ has a counteracting up-jump. This matching is always possible by construction. This proves the theorem.

A typical sample function $\Pi(t)$ is given in Fig. 3. Note the finite set of departures from monotonicity occurring at critical values $t_c$, mentioned in the proof. These points are those at which general position is temporarily lost. If $X_1, X_2, \ldots, X_n$ are independently drawn according to some probability density,
then, with probability one, any given $t$ will not be such an exceptional point. This will be used in the proof of Theorem 2.

The geometric result immediately implies that the probability of linear separability when the underlying distributions are shifted versions of one another is a nondecreasing function of the shift parameter. This probability is given by

$$P_t(n, d) = \int \frac{\Pi(t; x_1, x_2, \ldots, x_n)}{2^n} \prod_{i=1}^{n} f(x_i) dx_i$$

(8)

where $\Pi(t; x_1, x_2, \ldots, x_n)$ is the count function $\Pi(t)$ corresponding to $(x_1, x_2, \ldots, x_n)$ in the geometric theorem. Monotonicity can be shown by examining

$$P_t(n, d) - P_{t-\delta}(n, d) = \frac{\Pi(t; x_1, x_2, \ldots, x_n) - \Pi(t-\delta; x_1, x_2, \ldots, x_n)}{2^n} \prod_{i=1}^{n} f(x_i) dx_i$$

(9)

The proof of this fact is as follows.

The probability that $n$ points drawn at random from two sets having distribution densities $f_0(x)$ and $f_1(x)$ are separable about a fixed threshold $s$, on the line, is easily obtained as

$$P(n, 1; s) = \frac{1}{2^n} \left[ (1 + \psi(s))^{n-1} + (1 - \psi(s))^{n-1} \right]$$

(11)

where

$$\psi(s) = \int_{-\infty}^{s} \left( f_1(\xi) - f_0(\xi) \right) d\xi.$$

Now, note that we have a separable realization if and only if one of the points, say $x_i$, belongs to either class $\theta_i = 1$ or $\theta_i = -1$ and the others form a separable realization with $x_i$ as threshold. This proves (10), since separability is implied by the occurrence of one of $n$ disjoint events with probabilities given by either $\rho_+$ or $\rho_-$, where

$$\rho_\pm = \Pr[\theta_i = \pm 1] \int_{-\infty}^{s} P(n-1, 1; s) \Pr[X_i \in (s, s+ds) \mid \theta_i = \pm 1].$$

(13)

Although (10) seems to be asymmetric with respect to the class distributions, it is not difficult to recognize that the result is the same if we interchange $f_1(x)$ and $f_2(x)$.

Using expression (10), we readily prove the stated conjecture for $d = 1$. Indeed, if $f_1(x) \neq f_0(x)$, we always have $P(n, 1) >$
$P_0(n, 1) = n/2^{n-1}$, since the strict inequality $(1 + \psi(s))^{n-1} + (1 - \psi(s))^{n-1} > 2$ must hold for some values of $s$.

REFERENCES
