Discrete Mathematics 33 (1981) 313–318
North-Holland Publishing Company

SOME SPECIAL VAPNIK–CHERVONENKIS CLASSES

R.S. WENOCUR
Drexel University, Philadelphia, PA 19104, USA

R.M. DUDLEY*
Massachusetts Institute of Technology Cambridge, MA 02139, USA

Received 10 May 1980

For a class $\mathscr{C}$ of subsets of a set X, let $V(\mathscr{C})$ be the smallest n such that no n-element set $F \subset X$ has all its subsets of the form $A \cap F$, $A \in \mathscr{C}$. The condition $V(\mathscr{C}) < +\infty$ has probabilistic implications. If any two-element subset A of X satisfies both $A \cap C = \emptyset$ and $A \subset D$ for some $C, D \in \mathscr{C}$, then $V(\mathscr{C}) = 2$ if and only if $\mathscr{C}$ is linearly ordered by inclusion. If $\mathscr{C}$ is of the form $\mathscr{C} = \{\bigcap_{i=1}^{n} C_i : C_i \in \mathscr{C}_i,\ i = 1, 2, \ldots, n\}$, where each $\mathscr{C}_i$ is linearly ordered by inclusion, then $V(\mathscr{C}) \leq n + 1$. If H is an $(n-1)$-dimensional affine hyperplane in an n-dimensional vector space of real functions on X, and $\mathscr{C}$ is the collection of all sets $\{x : f(x) > 0\}$ for f in H, then $V(\mathscr{C}) = n$.

1. Introduction

Given a set X, a collection $\mathscr{C}$ of subsets of X, and a finite set $F \subset X$, let $\Delta^{\mathscr{C}}(F)$ denote the number of distinct sets $C \cap F$ for $C \in \mathscr{C}$. Define

$$m^{\mathscr{C}}(n) := \max(\Delta^{\mathscr{C}}(F) : \operatorname{card} F = n),$$

where, for finite F, $\operatorname{card} F$ is the number of elements in F. Let

$$V(\mathscr{C}) := \begin{cases} \inf\{n : m^{\mathscr{C}}(n) < 2^n\}, \\ +\infty \quad \text{if } m^{\mathscr{C}}(n) = 2^n \text{ for all } n. \end{cases}$$

Vapnik and Chervonenkis [8] introduced the above-mentioned quantities and showed that $m^{\mathscr{C}}(n) \leq n^{V(\mathscr{C})} + 1$.

Let $X_1, X_2, \ldots$ be i.i.d. (P), where P is a probability law on X. If $V(\mathscr{C}) < +\infty$, Vapnik and Chervonenkis determined that under suitable conditions, an unknown P can be approximated, uniformly on $\mathscr{C}$, by the empirical measure P_n. Furthermore, using empirical P_n and Q_m, it may be possible to distinguish between unknown laws P and Q; this is related to the problem of pattern recognition [8, 9]; for further probabilistic consequences, see [2, 4, 7].

For any probability space $(X, \mathscr{S}, P)$, let W_P be the Gaussian process, indexed by $\mathscr{S}$, with mean 0 and for all $A, B \in \mathscr{S}$, $EW_P(A)W_P(B) = P(A \cap B)$; W_P, of course, has independent values on disjoint sets. If, for each A, $G_P(A) := W_P(A) - P(A)W_P(X)$, G_P is a Gaussian process, indexed by $\mathscr{S}$, with zero mean, such that

* Partially supported by National Science Foundation Grant MCS-79-04474.

$EG_P(A)G_P(B) = P(A \cap B) - P(A)P(B)$ for all $A, B \in \mathcal{S}$. In [2], for a probability space $(X, \mathcal{S}, P)$, a collection $\mathcal{C} \subset \mathcal{S}$ is called a *P-Donsker class* if, together with suitable measurability conditions, the convergence of laws $\mathcal{L}(n^{\frac{1}{2}}(P_n - P)) \to \mathcal{L}(G_P)$ holds with respect to uniform convergence on $\mathcal{C}$ in an appropriate sense. In particular, a P-Donsker class must satisfy the property that G_P has a version that is almost surely bounded on $\mathcal{C}$. If this holds for all P on $\mathcal{S} \supset \mathcal{C}$, then $V(\mathcal{C}) < +\infty$ [4].

If $V(\mathcal{C}) < +\infty$ and $\mathcal{C}$ satisfies some measurability conditions, then $\mathcal{C}$ is P-Donsker for all P on $\mathcal{S} \supset \mathcal{C}$ [2, Section 7]. The following example [4] shows that $V(\mathcal{C}) < +\infty$ alone is not sufficient to give limit theorems uniformly on $\mathcal{C}$: consider the unit interval $I = [0, 1] \subset \mathbb{R}$. Assuming the continuum hypothesis, $[0, 1]$ can be well ordered by a relation $\prec$ so that for every $y \in I$, $\{x : x \prec y\}$ is countable. For $\mathcal{C} := \{\{x : x \prec y\};\ y \in I\}$, we have $V(\mathcal{C}) = 2$, yet if P is Lebesgue measure, then for all n, $\sup_{A \in \mathcal{C}} |(P_n - P)(A)| = 1$.

If $V(\mathcal{C}) < +\infty$, $\mathcal{C}$ is called a *Vapnik–Chervonenkis class* (VCC), and $V(\mathcal{C})$ its *Vapnik–Chervonenkis number* (VCN). Often, in applications, it is important to know what the sample size must be in order to assert, with a given high probability, that relative frequencies differ from their corresponding probabilities by a sufficiently small value over an entire class $\mathcal{C}$ of events. As a consequence of their main results, Vapnik and Chervonenkis [8] estimated, for VCC's, sample size by means of a function of the VCN. The study of special VCC's has this approach as an underlying motivation.

Some special VCC's have been investigated and their VCN's determined. For example, Vapnik and Chervonenkis [8] noted that for the class $\mathcal{C}_R$ of all rays $(-\infty, t]$ on the real line, $V(\mathcal{C}_R) = 2$. Results of Harding [6] imply that for the class $\mathcal{H}(k)$ of all halfspaces in k-dimensional Euclidean space, one has $V(\mathcal{H}(k)) = k + 2$. For the collection $\bar{B}(k)$ of all closed balls in $\mathbb{R}^k$, $V(\bar{B}(k)) = k + 2$ [3]; an easier proof appears below in Section 3. For general methods of generating VCC's, see also [2, Theorem 7.2, Proposition 7.12]. Here, we shall first examine VCC's of the form

$$\mathcal{C} = \left\{ \bigcap_{i=1}^{n} C_i : C_i \in \mathcal{C}_i,\ i = 1, 2, \ldots, n \right\},$$

where each $\mathcal{C}_i$ is a class of subsets of X linearly ordered by inclusion; then, in Section 3, we shall determine the VCN's for VCC's generated by certain collections of functions.

2. Classes linearly ordered by inclusion, and intersections

From the definition, it follows immediately that $V(\mathcal{C}) = 0$ iff $\mathcal{C}$ is empty, and that $V(\mathcal{C}) = 1$ iff $\mathcal{C}$ contains exactly one element. For the case $V(\mathcal{C}) = 2$, we begin with the following:

Definition. A finite set $F \subset X$ is *shattered* by $\mathscr{C}$ iff $2^F = \{F \cap C : C \in \mathscr{C}\}$.

Theorem 2.1. *Let* $\mathscr{C} = \{\bigcap_{i=1}^n C_i : C_i \in \mathscr{C}_i, i = 1, 2, \ldots, n\}$, *where each* $\mathscr{C}_i$ *is a collection of subsets of* X *that is linearly ordered by inclusion. Then* $V(\mathscr{C}) \leq n+1$.

Proof. Consider a set F of $n+1$ elements of X. For each i, $i = 1, 2, \ldots, n$, no more than one n-point subset of F is the intersection of F with an element of $\mathscr{C}_i$. Let A_i denote this n-point subset whenever it exists. If A is an n-point subset of F that satisfies $A = F \cap C$ for some $C = \bigcap_{i=1}^n C_i \in \mathscr{C}$, $C_i \in \mathscr{C}_i$, then for $j = 1, 2, \ldots, n$, either $C_j \cap F = A$ or $C_j \cap F = F$, with $C_i \cap F = A$ for at least one i, $1 \leq i \leq n$. Therefore, $A = A_i$ for some $i = 1, 2, \ldots, n$, which demonstrates that at most n n-point subsets of F can be realized as intersections of F with elements of $\mathscr{C}$. In conclusion, F is not shattered by $\mathscr{C}$.

Corollary 2.2. *If* $\mathscr{C}$ *contains at least two elements and is linearly ordered by inclusion, then* $V(\mathscr{C}) = 2$.

Proof. Since $\mathscr{C}$ contains at least two elements, $V(\mathscr{C}) \geq 2$, and the result follows from Theorem 2.1.

In $\mathbb{R}^n$, let $\mathscr{C}_i$ be the collection of all sets $\{x : x_i \leq a\}$, $a \in \mathbb{R}$, where x_i is the ith coordinate. Define $\mathscr{C}$ as in Theorem 2.1 and let F be the standard basis of unit vectors in $\mathbb{R}^n$. Then $\mathscr{C}$ shatters F, so Theorem 2.1 is sharp for all n. To restrict probability measures to $\mathscr{C}$ is, of course, to consider their n-dimensional distribution functions.

Also in $\mathbb{R}^n$, let $\mathscr{C}_{2i-1}$ be the collection of all sets $\{x : x_i \leq b\}$ and $\mathscr{C}_{2i}$ the collection of all sets $\{x : x_i \geq a\}$, where $a, b \in \mathbb{R}$, $i = 1, 2, \ldots, n$. Let

$$\mathscr{C} := \left\{ \bigcap_{i=1}^{2n} C_i : C_i \in \mathscr{C}_i \right\}.$$

Then by Theorem 2.1, $V(\mathscr{C}) \leq 2n+1$. On the other hand, let $G := \{e_1, -e_1, \ldots, e_n, -e_n\}$, where the e_i are standard basis vectors for $\mathbb{R}^n$. Then it is easy to see that $\mathscr{C}$ shatters G. We have thus found $V(\mathscr{C})$ for the class of rectangular sets:

Proposition 2.3. *Let*

$$\mathscr{C} := \left\{ \prod_{i=1}^n [a_i, b_i] : -\infty \leq a_i \leq b_i \leq +\infty, i = 1, 2, \ldots, n \right\}.$$

Then $V(\mathscr{C}) = 2n+1$. *If we require* $a_i = -\infty$ *for all* i, *then* $V(\mathscr{C}) = n+1$.

The following example shows that without further assumptions, the converse of Corollary 2.2 is not true. Let X be any set with at least 2 elements; let

$\mathscr{C} = \{\{x\} : x \in X\}$. Although $\mathscr{C}$ is not linearly ordered by inclusion, $V(\mathscr{C}) = 2$. A sufficiently strong hypothesis is provided by

Theorem 2.4. *Let $\mathscr{C}$ be a collection of subsets of X (with* card $X \geq 2$*) such that for every subset $A \subset X$ of 2 elements, there exist $U \in \mathscr{C}$ such that $A \subset U$, and $D \in \mathscr{C}$ such that $A \cap D = \emptyset$. Then $V(\mathscr{C}) = 2$ iff $\mathscr{C}$ is linearly ordered by inclusion.*

Proof. In light of Corollary 2.2, we need prove only one direction of our assertion. Suppose $\mathscr{C}$ is not linearly ordered by inclusion. Then there exist $B, C \in \mathscr{C}$ such that $B \not\subset C$ and $C \not\subset B$. Choosing $b \in B \setminus C$ and $c \in C \setminus B$, let $A = \{b, c\}$. Then A is shattered by $\mathscr{C}$, which implies $V(\mathscr{C}) > 2$.

Corollary 2.5. *If $\mathscr{C}$ is a collection of subsets of X such that $\{\emptyset, X\} \subset \mathscr{C}$, then $V(\mathscr{C}) = 2$ iff $\mathscr{C}$ is linearly ordered by inclusion.*

The collection of all rays $(-\infty, t]$ on the real line, for $-\infty < t < +\infty$, is an example of a class of sets that meets the requirements of Theorem 2.4, but fails to satisfy the hypothesis of its corollary. If we take $X = \{1, 2, \ldots\}$ and $\mathscr{C}$ as the collection of all finite initial segments $\{1, 2, \ldots, n\}$, $n \geq 0$, then $V(\mathscr{C}) = 2$; Theorem 2.4 applies, although Corollary 2.5 does not.

If $\mathscr{C}$ is as defined in Theorem 2.1, with $n = 2$, but is not linearly ordered by inclusion, and satisfies the hypothesis of Theorem 2.4, then $V(\mathscr{C}) = 3$. An example of such a class of sets is the collection of all extended intervals $(-\infty, t]$ in $\mathbb{R}^2$; another is the class of all closed intervals on the real line. As another special case, if

$$\mathscr{C} = \{L \cap M : L \in \mathscr{L}, M \in \mathscr{M}\} \cup \{\emptyset, X\},$$

where $\mathscr{L}$ and $\mathscr{M}$ are, but $\mathscr{C}$ itself is not, linearly ordered by inclusion, then $V(\mathscr{C}) = 3$.

3. Positivity sets for affine hyperplanes of real functions

If G is a collection of real-valued functions on a set X, define $\mathrm{pos}(G)$ to be the collection of all sets

$$\mathrm{pos}(g) := \{x \in X : g(x) > 0\}, \quad g \in G.$$

It is known that if G is an n-dimensional vector space of real functions on X, then $V(\mathrm{pos}(G)) = n + 1$ (Cover [1]; [2, Theorem 7.2]). Taking $X = \mathbb{R}^k$ and G as the space of affine functions, we reobtain that if $\mathscr{H}(k)$ denotes the set of (open) half-spaces of $\mathbb{R}^k$, then $V(\mathscr{H}(k)) = k + 2$ for all $k = 1, 2, \ldots$. Now, let H be an m-dimensional real vector space of real functions on X, $f \notin H$ a real function on

X, and define

$$H_1 := \{f + h : h \in H\},$$
$$G := \{\alpha f + h : h \in H, \alpha \in \mathbb{R}\}.$$

Since $H_1 \subset G$, it follows that $V(\text{pos}(H_1)) \leq m+2$. Actually, however, we have

Theorem 3.1. *If H is an m-dimensional real vector space of real functions on a set X, $f \notin H$ a real function on X, and $H_1 := \{f + h, h \in H\}$, then $V(\text{pos}(H_1)) = m+1$.*

Proof. The argument is similar to that in [2, proof of Theorem 7.2]. Let $A \subset X$ with card $A = m+1$. G as defined above is an $m+1$-dimensional real vector space of functions. The map $r_A : G \to \mathbb{R}^A$ that restricts functions in G to the set A is a linear map, which may fail to be surjective. If this is the case, we can find nonzero v in $\mathbb{R}^A$ orthogonal to $r_A(G)$, in particular, to $r_A(H_1)$, with respect to $(\cdot, \cdot)_A$, the usual inner product on $\mathbb{R}^A$. If $A_+ := \{x \in A : v(x) > 0\}$, we may assume that A_+ is nonempty, since otherwise we can select $-v$ (as M. Artin noted). Suppose $A_+ = A \cap C$, $C \in \text{pos}(G)$. Then there exists $g \in G$ with $\{x \in X : g(x) > 0\} = C$. This implies $(r_A(g), v) > 0$, a contradiction. Now, suppose r_A is onto. Viewing G and $\mathbb{R}^A$ as affine spaces, where we identify "points" with their vector representations, H_1 is a hyperplane in G. It follows that $r_A(H_1)$ is a hyperplane in $\mathbb{R}^A$ that does not include 0. Hence, there exists $v \in \mathbb{R}^A$ such that $(h, v)_A = -1$ for all $h \in r_A(H_1)$. Again, let $A_+ = \{x \in A : v(x) > 0\}$. A_+ may be empty. If $A_+ = A \cap C$, $C \in \text{pos}(H_1)$, take $g \in H_1$ such that $C = \{x \in X : g(x) > 0\}$. We have $(r_A(g), v)_A \geq 0$, a contradiction. Therefore, $V(\text{pos}(H_1)) \leq m+1$. To show that $V(\text{pos}(H_1)) > m$, we appeal to the property that H_1 is an m-variety of G. This implies that there exists a translation τ such that $H = H_1 \tau$ and a subset B of X with card $B = m$ such that $r_B(H) = \mathbb{R}^B$, so all subsets of B are of the form $B \cap C$, $C \in \text{pos}(H_1)$.

As an example of a situation in which $A_+ = \emptyset$, let $X = \mathbb{R}^2 = \{(x, y)\}$; let H be the vector space spanned by the first coordinate function, x; let $f = y$. Select $A = \{a_1, a_2\}$ where $a_1 = (-1, -1)$; $a_2 = (1, 2)$. Represent an element of $\mathbb{R}^A$ as $\langle \gamma_1, \gamma_2 \rangle$, where γ_1 is the a_1st-coordinate and γ_2 is the a_2nd-coordinate. Then $r_A(H_1) = \{\langle \alpha, 1-\alpha \rangle : \alpha \in \mathbb{R}\}$, and $v = \langle -1, -1 \rangle$. Therefore, $A_+ = \emptyset$; we cannot achieve $\emptyset$ by intersections of A with elements of $\text{pos}(H_1)$. We can see this graphically as well, since $\text{pos}(H_1)$ consists of the "top halves" of $\mathbb{R}^2$ determined by all lines, except $x = 0$, through the origin. This also provides an example of a collection $\mathcal{C}$ that is not linearly ordered by inclusion, yet satisfies $V(\mathcal{C}) = 2$. Another (simpler) such example appears in Section 2 above.

For a collection G of functions on a set X, we define $\text{nn}(G)$ to be the collection of all sets

$$\text{nn}(g) := \{x \in X : g(x) \geq 0\}, \quad g \in G.$$

With H, f, and H_1 as defined in Theorem 3.1, let $H_2 := \{-f + h : h \in H\}$. By

Theorem 3.1, $V(\mathrm{pos}(H_2)) = m+1$. Taking complements of elements of $\mathrm{pos}(H_2)$, we have

Corollary 3.2. $V(\mathrm{nn}(H_1)) = m+1$.

If H on $\mathbb{R}^k$ is the vector space spanned by 1 and the coordinates $x_1, x_2, \ldots, x_k$, and if $f = -|x|^2$ for $x \in \mathbb{R}^k$, where $|\cdot|$ denotes the usual Euclidean norm, we obtain from Theorem 3.1 and Corollary 3.2:

Corollary 3.3. *If $B(k)$ denotes the set of all open balls*

$$B(x, s) := \{y : |x-y| < s\}, \quad x \in \mathbb{R}^k, s > 0,$$

and if $\bar{B}(k)$ denotes the set of all closed balls

$$\bar{B}(x, s) := \{y : |x-y| \leqslant s\}, \quad x \in \mathbb{R}^k, s > 0,$$

then $V(B(k)) = V(\bar{B}(k)) = k+2$ for all $k = 1, 2, \ldots$.

The result for closed balls appeared in [3], with a longer proof.

References

[1] T.M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications to pattern recognition, IEEE Trans. Electron. Comput. EC-14 (1965) 326–334.

[2] R.M. Dudley, Central limit theorems for empirical measures, Ann. Probability 6 (1978) 899–929; Correction, 7 (1979) 909–911.

[3] R.M. Dudley, Balls in $\mathbb{R}^k$ do not cut all subsets of $k+2$ points, Adv. in Math. 31 (3) (1979) 306–308.

[4] M. Durst and R.M. Dudley, Empirical processes, Vapnik–Chervonenkis classes and Poisson processes, in: K. Urbanik, ed., Probability and Mathematical Statistics 1, Wrocław, Poland (1980), to appear.

[5] P. Gaenssler and W. Stute, Empirical processes: a survey of results for independent and identically distributed random variables, Ann. Probability 7 (2) (1979) 193–243.

[6] E.F. Harding, The number of partitions of a set of N points in k dimensions induced by hyperplanes, Proc. Edinburgh Math Soc. 15 (1967) 285–289.

[7] J. Kuelbs and R.M. Dudley, Log log laws for empirical measures, Ann. Probability 8 (1980) 405–418.

[8] V.N. Vapnik and A.Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theor. Probability Appl. 16 (1971) 264–280 (English); Teor. Verojatnost. i Primenen. 16 (1971) 264–279 (Russian).

[9] V.N. Vapnik and A.Ya. Chervonenkis, Theory of Pattern Recognition (in Russian) (Nauka, Moscow, 1974).