VC-dimension

Let $\mathcal{F} \subseteq \mathcal{P}(X)$ and $S \subseteq X$. The *trace of* $\mathcal{F}$ *on* S is the set system

$$\mathcal{F}|S := \{A \cap S : \text{there exists } F \in \mathcal{F} \text{ such that } F \cap S = A\}.$$

We set

$$\mathrm{tr}_{\mathcal{F}}(S) = |\mathcal{F}|S|\,,$$

i.e. the number of sets in the trace of $\mathcal{F}$ on S. We say that S *is shattered by* $\mathcal{F}$ if $\mathcal{F}|S = \mathcal{P}(S)$ (in other words, $\mathrm{tr}_{\mathcal{F}}(S) = 2^{|S|}$).

The *VC-dimension of* $\mathcal{F}$ is $\max\{|S| : S \subseteq X \text{ is shattered by } \mathcal{F}\}$. [VC stands for Vapnik-Chervonenkis.]

Example 1. The family $[n]^{\leq d}$ has VC-dimension d.

What is the VC-dimension of the (infinite) family $\mathcal{H}$ consisting of all half-planes (in $\mathbb{R}^2$)? For instance, $\{(1,1),(2,1),(3,1)\}$ cannot be shattered by $\mathcal{H}$ (there is not way to obtain the subset $\{(1,1),(3,1)\}$ by intersecting $\{(1,1),(2,1),(3,1)\}$ with half-planes!). However, $\{(0,1),(1,0),(1,2)\}$ is easily seen to be shattered by $\mathcal{H}$. So the VC-dimension of $\mathcal{H}$ is at least 3. (Tricky question: What is the VC-dimension of this system?)

The Sauer-Shelah Theorem tells us that if a family $\mathcal{A} \subseteq \mathcal{P}(n)$ contains more than $|[n]^{(\leq d)}|$ sets, then its VC-dimension is greater than d:

Theorem 21. *If* $\mathcal{A} \subseteq \mathcal{P}(n)$ *has VC-dimension at most* d*, then*

$$|\mathcal{A}| \leq |[n]^{(\leq d)}| = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}.$$

We shall see a couple of proofs.

Proof 1 of the Sauer-Shelah Theorem. We argue by induction on $n + d$. Let

$$f(n,d) = |[n]^{(\leq d)}| = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}.$$

If $n = 0$ or $d = 0$, the result is trivial. If $n + d > 0$, with $n, d > 0$, let

$$\mathcal{B} = \{A \setminus n : A \in \mathcal{A}\} \subseteq \mathcal{P}(n-1),$$

$$\mathcal{C} = \{A \in \mathcal{A} : n \notin A, A \cup \{n\} \in \mathcal{A}\} \subseteq \mathcal{P}(n-1).$$

Then $\mathcal{B}$ has VC-dimension at most d, while $\mathcal{C}$ has VC-dimension at most $d-1$ (as if S is shattered by $\mathcal{C}$, then $S \cup \{n\}$ is shattered by $\mathcal{A}$; so $|S| \leq d-1$). Hence, by induction, $|\mathcal{B}| \leq f(n-1, d)$ and $|\mathcal{C}| \leq f(n-1, d-1)$ and

$$|\mathcal{A}| = |\mathcal{B}| + |\mathcal{C}| \leq f(n-1, d) + f(n-1, d-1) = f(n, d).$$

■

Proof 2 of the Sauer-Shelah Theorem. Define the *i-compression operator* by

$$\pi_i(A) = A \setminus \{i\}$$

and

$$\pi_i(\mathcal{A}) = \{\pi_i(A) : A \in \mathcal{A}\} \cup \{A \in \mathcal{A} : \pi_i(A) \in \mathcal{A}\}.$$

Then π_i does not increase the VC-dimension of a set system (exercise) and $|\mathcal{A}| = |\pi_i(\mathcal{A})|$. Thus we can repeatedly apply the i-compression operator until our family is i-compressed for all $i \in [n]$ (this terminates, as every compression either leaves the family unchanged or decreases the quantity $\sum_{A \in \mathcal{A}} |A|$.

So consider $\mathcal{B}$, the i-compressed family obtained from $\mathcal{A}$. If $\mathcal{B}$ contains any set B of size at least $d+1$ then $\mathcal{B}$ contains all subsets of B, and so has VC-dimension at least $d+1$. Otherwise,

$$|\mathcal{A}| = |\mathcal{B}| \leq |[n]^{(\leq d)}| = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}.$$

A brief interlude on upsets and downsets

A family $\mathcal{A}$ is an *upset* if $A \in \mathcal{A}$ and $A \subseteq B$ implies that $B \in \mathcal{A}$. $\mathcal{A}$ is a *downset* if $A \in \mathcal{A}$ and $A \supset B$ implies that $B \in \mathcal{A}$.

Theorem 22. (Kleitman's Theorem) *Let $\mathcal{A}, \mathcal{B} \subseteq \mathcal{P}(n)$ be downsets. Then*

$$|\mathcal{A} \cap \mathcal{B}| \geq \frac{|\mathcal{A}||\mathcal{B}|}{2^n}.$$

Proof. We argue by induction on n. The case $n = 1$ is straightforward. For $n > 1$, define

$$\mathcal{A}^+ = \{A \subseteq [n-1] : A \cup \{n\} \in \mathcal{A}\}$$

and

$$\mathcal{A}^- = \{A \subseteq [n-1] : A \in \mathcal{A}\}.$$

Define $\mathcal{B}^+$ and $\mathcal{B}^-$ similarly.

Since $\mathcal{A}$ is a downset, $\mathcal{A}^+, \mathcal{A}^-$ are downsets and $\mathcal{A}^+ \subseteq \mathcal{A}^-$; similarly for $\mathcal{B}^+$ and $\mathcal{B}^-$. Then, by induction,

$$\begin{aligned}
|\mathcal{A} \cap \mathcal{B}| &= |\mathcal{A}^+ \cap \mathcal{B}^+| + |\mathcal{A}^- \cap \mathcal{B}^-| \\
&\geq \frac{|\mathcal{A}^+||\mathcal{B}^+|}{2^{n-1}} + \frac{|\mathcal{A}^-||\mathcal{B}^-|}{2^{n-1}} \\
&= \frac{1}{2^n}(|\mathcal{A}^+| + |\mathcal{A}^-|)(|\mathcal{B}^+| + |\mathcal{B}^-|) + \frac{1}{2^n}(|\mathcal{A}^+| - |\mathcal{A}^-|)(|\mathcal{B}^+| - |\mathcal{B}^-|) \\
&\geq \frac{|\mathcal{A}||\mathcal{B}|}{2^n},
\end{aligned}$$

since $(|\mathcal{A}^+| - |\mathcal{A}^-|) \leq 0$ and $(|\mathcal{B}^+| - |\mathcal{B}^-|) \leq 0$. ■

Some authors call the above Theorem "Harris' Lemma" or "Harris-Kleitman Lemma".