L20

Recall **VC-dimension** of a range space (X, R) is the size of the largest subset $A \subseteq X$ that can be **shattered** by R (or ∞ if no largest exists).

$\{A \cap r \mid r \in R\} = 2^A$

Intuition Range spaces with finite VC-dimension are "simpler" than ones with infinite VC-dim.

the projection of R to A

Let $R|_A \doteq \{r \cap A \mid r \in R\}$ $\longrightarrow$ $N \subseteq A$ is an ε-net for A if for all $r \in R|_A$ $|r| > \varepsilon |A| \Rightarrow r \cap N \neq \emptyset$

Sauer's Lemma (Shelah's) If (X, R) is a range space with finite VC-dim at most d then

$$|R|_A| \leq \phi_d(n) \equiv \sum_{i=0}^{d} \binom{n}{i}$$

$O(n^d)$ not 2^n

for all $A \subseteq X$ with $|A| = n$

---

old proof First observe

$$\phi_d(n) = \begin{cases} 0 & \text{if } d = -1 \\ 1 & \text{if } n = 0 \\ \phi_d(n-1) + \phi_{d-1}(n-1) & \text{o.w.} \end{cases}$$

$\binom{n}{0} = 1$

$\binom{n}{k} = \underset{\substack{\uparrow \\ \text{don't take first}}}{\binom{n-1}{k}} + \underset{\substack{\uparrow \\ \text{do take first}}}{\binom{n-1}{k-1}}$

Second

$VC(A, R|_A) \leq VC(X, R) \leq d$

Consider $(A \setminus \{x\}, R - x)$ and $(A \setminus \{x\}, R^{(x)})$

$R - x = \{r \setminus \{x\} \mid r \in R|_A\}$

$R^{(x)} = \{r \in R|_A \mid x \notin r \ \&\ r \cup \{x\} \in R|_A\}$

Note: $\boxed{|R|_A| = |R - x| + |R^{(x)}|}$

- By induction $\boxed{|R - x| \leq \phi_d(n-1)}$
- If $S \subseteq A \setminus \{x\}$ is shattered by $R^{(x)}$ then $S \cup \{x\}$ is shattered by $R|_A$

$\Rightarrow VC(A \setminus \{x\}, R^{(x)}) \leq d - 1$ and $\boxed{|R^{(x)}| \leq \phi_{d-1}(n-1)}$

Pajor's Theorem

Every finite set family F shatters at least $|F|$ sets

proof (by induction)

base: Every F with $|F| = 1$ shatters the empty set

Step: If $|F| > 1$, let x be an element in some but not all sets in F

Let $F_x = \{ S \in F \mid x \in S \}$ and $F_{\bar{x}} = \{ S \in F \mid x \notin S \}$

by induction F_x shatters at least $|F_x|$ sets

and $F_{\bar{x}}$ shatters at least $|F_{\bar{x}}|$ sets with $|F| = |F_x| + |F_{\bar{x}}|$

are we done? NO overlap

None of the sets shattered by F_x or $F_{\bar{x}}$ contain x.

If S is shattered by both F_x and $F_{\bar{x}}$ then both S and $S \cup \{x\}$ are shattered by F

If S is shattered by one of F_x or $F_{\bar{x}}$ then S is shattered by F.

$\Rightarrow$ F shatters at least $|F_x| + |F_{\bar{x}}| = |F|$ sets. $\square$

Example

$F = \{a,b\} \{a,c\}, \{c\}$ $[\emptyset, \{a\}, \{b\}]$

$x = a$

$\{a,b\} \{a,c\}$ $\{c\}$ $[\emptyset]$

$x = b$ $[\emptyset, \{b\}]$

$\{a,b\}$ $\{a,c\}$

$[\emptyset]$ $[\emptyset]$

Pajor $\Rightarrow$ Sauer/Shelah F

If $|R|_A| > \Phi_d(n) = \sum_{i=0}^{d} \binom{n}{i}$ then, by Pajor,

there are more than $\Phi_d(n)$ sets shattered by $R|_A$

but there are only $\Phi_d(n)$ subsets of A of size $\leq d$

so $R|_A$ must shatter some subset of A of size $> d$

$\Rightarrow$ VC-dim $(X, R) > d$ $\Rightarrow\Leftarrow$

How to find an ε-net for a set $A \subseteq X$ w.r.t. range space (X, R)

Thm If $VC(X,R) \le d$ and $\varepsilon \le \frac{1}{2}$ then there exists an ε-net for $(A, R|_A)$ w.r.t. μ of size at most $\frac{cd}{\varepsilon} \ln \frac{1}{\varepsilon}$ for some constant c.

a probability measure μ on A

≤ 20 or even less for smaller ε

proof Let $\boxed{s = \frac{cd}{\varepsilon} \ln 1/\varepsilon}$ ← assume this is an integer

Let N be a random sample of size s drawn indep. w.r.t. μ from A.

To show: N is ε-net with at least constant probability.

Let $F \equiv R|_A$

Assume all $S \in F$ have $\mu(S) \ge \varepsilon$ (smaller sets don't matter)

For each $S \in F$: $\Pr[N \cap S = \emptyset] \le (1-\varepsilon)^s \le e^{-\varepsilon s}$

So if $s \ge \frac{1}{\varepsilon} \ln(|F|+1)$ we're done

BUT $|F|$ may be $\Omega(n^d)$ ← VC-dim

too big

We need a tool and a trick

Tool Let $X = X_1 + X_2 + \dots + X_n$ where X_i are indep random variables with $\Pr[X_i = 1] = p$, $\Pr[X_i = 0] = 1-p$

Then $\Pr[X \ge \frac{1}{2} np] \ge \frac{1}{2}$ (when $np \ge 8$)

proof ⟨Chebyshev's Ineq.⟩

$$\Pr[|X - E[X]| < t] \le \frac{Var[X]}{t^2}$$

$E[X] = np \qquad Var[X] = \sum_i Var[X_i] \le np$

$$\Pr[X < \tfrac{1}{2} np] \le \Pr[|X - E[X]| \ge \tfrac{1}{2} np] \le \tfrac{4}{np} \le \tfrac{1}{2} \text{ for } np \ge 8 \quad \square$$