CSCI4230 Computational Learning Theory — Spring 2021

Lecturer: Siu On Chan — *Based on Rocco Servedio's notes and Wikipedia*

Notes 13: Sauer–Shelah lemma

1. Sauer–Shelah Lemma

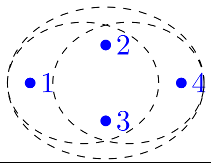
Claim 1. $|\Pi_{\mathcal{C}}(S)| \leqslant |\{T \subseteq S \mid \mathcal{C} \text{ shatters } T\}|$

Proof. Apply following Proposition with $\mathcal{F} = \Pi_{\mathcal{C}}(S)$

Note that T is shattered by $\mathcal{C}$ if and only if T is shattered by $\mathcal{F} = \Pi_{\mathcal{C}}(S)$ □

Proposition 2 (Pajor)**.** *A finite family $\mathcal{F}$ of subsets over S shatters at least $|\mathcal{F}|$ subsets, i.e.*

$$|\mathcal{F}| \leqslant \#\textit{subsets } \mathcal{F} \textit{ shatters} = |\{T \subseteq S \mid \mathcal{F} \textit{ shatters } T\}|$$

e.g. 

$$\mathcal{F} = \left\{ \begin{matrix} \{1,2,3\}, \\ \{2,3,4\}, \\ \{1,2,3,4\} \end{matrix} \right\}, \qquad \mathcal{F} \text{ shatters } \{1\}, \{4\}, \emptyset$$

Proof of Proposition. Base case $|\mathcal{F}| = 0$: trivial

Base case $|\mathcal{F}| = 1$: $\mathcal{F}$ shatters $\emptyset$

Induction step for $|\mathcal{F}| > 1$: Fix $x \in S$ belonging to some but not all of the sets in $\mathcal{F}$

Split $\mathcal{F}$ into $\mathcal{F}_{\ni x}$ and $\mathcal{F}_{\not\ni x}$ (those containing x and those do not)

Induction hypothesis implies $\mathcal{F}_{\ni x}$ shatters $\geqslant |\mathcal{F}_{\ni x}|$ subsets, $\mathcal{F}_{\not\ni x}$ shatters $\geqslant |\mathcal{F}_{\not\ni x}|$ subsets

$$|\mathcal{F}| = |\mathcal{F}_{\ni x}| + |\mathcal{F}_{\not\ni x}| \leqslant \#\text{subsets } \mathcal{F}_{\ni x} \text{ shatters} + \#\text{subsets } \mathcal{F}_{\not\ni x} \text{ shatters}$$

Remains to show right-hand-side $\leqslant$ #subsets $\mathcal{F}$ shatters

Any set shattered by $\mathcal{F}_{\ni x}$ cannot contain x, since all sets in $\mathcal{F}_{\ni x}$ contain x

Any set shattered by $\mathcal{F}_{\not\ni x}$ cannot contain x, since all sets in $\mathcal{F}_{\not\ni x}$ do not contain x

Thus any set of the form $T \cup \{x\}$ cannot be shattered by $\mathcal{F}_{\ni x}$ or $\mathcal{F}_{\not\ni x}$

If T is shattered by only one of $\mathcal{F}_{\ni x}$ or $\mathcal{F}_{\not\ni x}$, T contributes 1 to #subsets $\mathcal{F}$ shatters

If T is shattered by both $\mathcal{F}_{\ni x}$ and $\mathcal{F}_{\not\ni x}$, then T and $T \cup \{x\}$ are both shattered by $\mathcal{F}$

T and $T \cup \{x\}$ together contribute 2 to #subsets $\mathcal{F}$ shatters □

Lemma 3 (Perles–Sauer–Shelah)**.** *When* $\mathrm{VCDim}(\mathcal{C}) = d$*,* $\Pi_{\mathcal{C}}(m) \leqslant \binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{d}$

Proof. By above Claim, at most $\displaystyle\sum_{0 \leqslant k \leqslant d} \binom{m}{k}$ choices for shattered subset T

No subset larger than $d = \mathrm{VCDim}(\mathcal{C})$ is shattered □

Corollary 4. *When* $\mathrm{VCDim}(\mathcal{C}) = d$ *and* $m \geqslant d$*,* $\Pi_{\mathcal{C}}(m) \leqslant \left(\frac{em}{d}\right)^d$

Proof. Want to show $\displaystyle\sum_{0 \leqslant k \leqslant m} \binom{m}{k} \leqslant \left(\frac{em}{d}\right)^d$ for $m \geqslant d$

$$\left(\frac{d}{m}\right)^d \sum_{0 \leqslant k \leqslant d} \binom{m}{k} \leqslant \sum_{0 \leqslant k \leqslant d} \left(\frac{d}{m}\right)^k \binom{m}{k} \leqslant \sum_{0 \leqslant k \leqslant m} \left(\frac{d}{m}\right)^k \binom{m}{k} = \left(1 + \frac{d}{m}\right)^m \leqslant (e^{d/m})^m = e^d$$

First inequality due to $d/m \leqslant 1$

Second inequality due to $d \leqslant m$

Next equality is binomial theorem

Last inequality is $1 + x \leqslant e^x$ for all real x □

---

2. Consistent Hypothesis

Theorem 5. *Given m independent labelled samples, with prob. $\geqslant 1-\delta$, any hypothesis consistent with all m samples has erorr at most ε, provided*

$$m \geqslant \Omega\left(\frac{1}{\varepsilon}\log\frac{\Pi_{\mathcal{C}}(2m)}{\delta}\right)$$

Compared with notes09, now $\mathcal{C}$ may be infinite
notes09 was union bound over $\mathcal{H}$; now over dichotomies on $2m$ samples

Proof. Imagine drawing $2m$ labelled samples $(x^i, c(x^i))$ from $\mathrm{EX}(c,\mathcal{D})$
Call m of the samples S_1; the remaining m samples S_2
Event A: Some bad $h \in \mathcal{C}$ is consistent with S_1
    Recall h is bad if $\mathrm{err}_{\mathcal{D}}(h,c) \geqslant \varepsilon$;   Goal: show $\mathbb{P}[A] \leqslant \delta$
Event B: Some $h \in \mathcal{C}$ is consistent with S_1 but wrong on $\geqslant \varepsilon m/2$ samples in S_2

Claim 6. *If $m \geqslant 8/\varepsilon$, then $\mathbb{P}[A] \leqslant 2\,\mathbb{P}[B]$*

Proof of Claim. $\mathbb{P}[B] \geqslant \mathbb{P}[B \text{ and } A] = \mathbb{P}[A]\,\mathbb{P}[B \mid A]$
Suffice to show $\mathbb{P}[B \mid A] \geqslant 1/2$
When A occurs, fix any bad h, $\mathbb{P}[h \text{ makes at most } \varepsilon m/2 \text{ mistakes on } S_2] \leqslant e^{-\frac{1}{8}\varepsilon m} \leqslant 1/e \leqslant 1/2$ □

Using Claim, suffices to show $\mathbb{P}[B] \leqslant \delta/2$
Equivalent way to view B:

(1) First draw $2m$ independent labelled samples S
(2) Randomly split S into two halves, S_1 and S_2 (first and second halves)
(3) Event B: S_1 contains no mistakes, S_2 contains $\geqslant \varepsilon m/2$ mistakes

Now fix any $2m$ instances S and a labeling/dichotomy of S (from $\Pi_{\mathcal{C}}(S)$) from step (1)
Event B is equivalent to $\geqslant \varepsilon m/2$ mistakes in S all falling in S_2
Combinatorial experiment: $2m$ balls (S), each colored red (mistake) or blue (correct)
    exactly ℓ are red   ($\ell \geqslant \varepsilon m/2$)
    Randomly put m balls into S_1 and the other m balls into S_2
    $\mathbb{P}[\text{all red balls fall into } S_2 \text{ equals}] = \binom{m}{\ell}/\binom{2m}{\ell}$
    $(= \mathbb{P}[\text{out of } 2m \text{ uncolored balls, randomly color } \ell \text{ of them red and all red balls fall on } S_2])$

$$\frac{\binom{m}{\ell}}{\binom{2m}{\ell}} = \frac{m}{2m}\frac{m-1}{2m-1}\cdots\frac{m-\ell+1}{2m-\ell+1} \leqslant \left(\frac{1}{2}\right)^{\ell}$$

Union bound over at most $\Pi_{\mathcal{C}}(S)$ labelings of S with $\ell \geqslant \varepsilon m/2$:

$$\mathbb{P}[B] \leqslant \frac{\Pi_{\mathcal{C}}(2m)}{2^{\varepsilon m/2}} \leqslant \frac{\delta}{2} \qquad \text{when } m \geqslant \frac{2}{\varepsilon}\log\frac{2\Pi_{\mathcal{C}}(2m)}{\delta}$$

Advantage of Event B over Event A:
    union bound over finitely many (in fact $\Pi_{\mathcal{C}}(2m)$) labelings; even when $\mathcal{C}$ is infinite □