**CSE 513T Theory of Artificial Intelligence & Machine Learning** February 13, 2018

# Lecture 9

*Lecturer: Brendan Juba* *Scribe: Sirui Li*

## Contents

**1. Definitions from last two lectures** **1**

**2. Start sample complexity upper bound** **1**
- reducing to finite sample

**3. Sauer's Lemma** **2**

**4. Proof of sample complexity upper bound, part 1** **3**

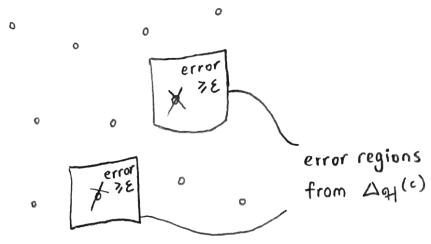
## 1 Definitions from last two lectures

**Dichotomies** For any class of representations $\mathcal{H}$ over $\mathcal{X}$, $\mathcal{S} \subseteq \mathcal{X}$, the set of dichotomies of $\mathcal{S}$ realized by $\mathcal{H}$
$\Pi_{\mathcal{H}}(\mathcal{S}) \stackrel{\text{def}}{=} \{(h(x_1), h(x_2), ..., h(x_m)) : h \in \mathcal{H}, \mathcal{S} = \{x_1, x_2, ..., x_m\}\}$

**Shattering** If $\Pi_{\mathcal{H}}(\mathcal{S}) = \{0,1\}^m$, we say $\mathcal{S}$ is shattered by $\mathcal{H}$.

**VC-dimension** The VC-dimension of $\mathcal{H}$ is the size of the largest set shattered by $\mathcal{H}$.

**Error Regions** For any target property c, the error regions of $\mathcal{H}$ w.r.t. c is
$\Delta_{\mathcal{H}}(c) \stackrel{\text{def}}{=} \{[h \neq c](x) : h \in \mathcal{H}\}$
The error regions with probability $\geq \epsilon$ under distribution $\mathcal{D}$ is $\Delta_{\mathcal{H},\epsilon}(c)$




**$\epsilon$-net** For any $\epsilon > 0$, the set $\mathcal{S} \subseteq \mathcal{X}$ is an $\epsilon$-net for $\Delta_{\mathcal{H}}(c)$ w.r.t. $\mathcal{D}$ if for every $r \in \Delta_{\mathcal{H},\epsilon}(c)$, there is an example $x \in \mathcal{S}$ s.t. $x \in r$

**Growth Function** The growth function of $\mathcal{H}$, $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ is $\Pi_{\mathcal{H}}(m) = \max_{\substack{\mathcal{S} \subseteq \mathcal{X}: \\ |\mathcal{S}|=m}} |\Pi_{\mathcal{H}}(\mathcal{S})|$

## 2 Sample complexity upper bound, goal

**Theorem:** Let $c$ be any property on $\mathcal{X}$, let $\mathcal{H}$ be any class of representations of VC-dimension $d$. Then any solution $h$ to Consis$_{\mathcal{H}}$ on a data set of $m \geq c_0 \cdot \frac{1}{\epsilon}(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta})$ examples drawn independently from a distribution $\mathcal{D}$ on $\mathcal{X}$, labeled by $c$ satisfies $\Pr_{x \in \mathcal{D}}[h(x) = c(x)] \geq 1 - \epsilon$ with probability $1 - \delta$.

**Strategy:** it suffices to show that with probability $1-\delta$, the dataset forms an "$\epsilon$-net" for the error regions of $\mathcal{H}$ w.r.t. $c$ ("$\Delta_{\mathcal{H}}(c)$").
We can't afford to take a union bound over all of $\mathcal{H}$ like in Occam's Razor ($\mathcal{H}$ may be infinite...), but since $\mathcal{H}$ has small VC-dimension, we can bound the growth function (# of possible labelings of a set of m examples) by Sauer's Lemma. So we'll be able to take a union bound over the # of possible labelings of the data on a small set of points.
Key: relating probability of getting $\epsilon$-net to the behavior on $\approx$ m points.

# 3 Sauer's Lemma

**Sauer's Lemma:** If $\mathcal{H}$ is a class of representations of VC-dimension $d$, then $\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} = O(m^d)$

**Proof:** For any $m, d \in \mathbb{N}$, define $\Phi_d(m)$ inductively by $\Phi_d(0) = \Phi_0(m) = 1$, and for $m, d \geq 0$, $\Phi_d(m) = \Phi_d(m-1) + \Phi_{d-1}(m-1)$

**Lemma 1:** If VC-dimension of $\mathcal{H}$ is $d$, then for every $m$, $\Pi_{\mathcal{H}}(m) \leq \Phi_d(m)$

**Lemma 2:** $\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$ (last time)

(For $m \leq d$, we saw $\Pi_{\mathcal{H}}(m) = 2^m = \sum_{i=0}^{d} \binom{m}{i}$)

For $m > d$, since $0 < \frac{d}{m} < 1$,

$$\left(\frac{d}{m}\right)^d \sum_{i=0}^{d} \binom{m}{i} \leq \sum_{i=0}^{d} \left(\frac{d}{m}\right)^i \binom{m}{i}$$

$$\boxed{\text{since } \left(\tfrac{d}{m}\right)^i \binom{m}{i} \geq 0 \ \ \forall i} \leq \sum_{i=0}^{m} \left(\frac{d}{m}\right)^i \binom{m}{i}$$

$$= \left(1 + \frac{d}{m}\right)^m$$

$$= \left(1 + \frac{d}{m}\right)^{\frac{m}{d} d}$$

$$\leq e^d$$

divide by $\left(\frac{d}{m}\right)^d$ ...

$$\sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{e \cdot m}{d}\right)^d = O(m^d) \quad \checkmark$$

**Proof of Lemma 1 by induction on m & d:**
**Base: m=0 or d=0.**
If $d = 0$, every $h \in \mathcal{H}$ must give the same values for all of $\mathcal{X}$, so $\Pi_{\mathcal{H}}(m) = 1 = \Phi_0(m)$.
Similarly, if $m = 0$, since we have no examples to label, $\Pi_{\mathcal{H}}(0) \leq 1 = \Phi_d(0)$.
**IH:** for $d', m'$ s.t. $d' \leq d, m' \leq m$, and at least one is strictly smaller, $\Pi_{\mathcal{H}}(m') \leq \Phi_{d'}(m')$.
**Ind. Step:**
Consider any set $\mathcal{S}$ of size $m$. We'll show $\Phi_d(m) \geq \Pi_{\mathcal{H}}(\mathcal{S})$ and therefore by def. $\Phi_d(m) \geq \Pi_{\mathcal{H}}(m)$.
Choose any $x \in \mathcal{S}$ (note $m \geq 1$).

Note $|\mathcal{S} - \{x\}| = m - 1$. So, $\Pi_{\mathcal{H}}(\mathcal{S} - \{x\}) \leq \Phi_d(m-1)$.
Note that $\Pi_{\mathcal{H}}(\mathcal{S} - \{x\})$ is $\Pi_{\mathcal{H}}(\mathcal{S})$ projected down to $\mathcal{S} - \{x\}$, so we can find $|\Pi_{\mathcal{H}}(\mathcal{S})|$ by taking $|\Pi_{\mathcal{H}}(\mathcal{S} - \{x\})|$ and adding back the # labelings that collide, i.e. that agree on $\mathcal{S} - \{x\}$, and differ on $x$.

Let $\mathcal{B}$ denote the set of $\mathbf{b} \in \Pi_{\mathcal{H}}(\mathcal{S} - \{x\})$ that could be reached by two different members of $\Pi_{\mathcal{H}}(\mathcal{S})$.
Thinking of $\mathbf{b}$ as a set ($b_{x_i} = 1 \Leftrightarrow x_i \in b$).
$|\mathcal{B}| = |\underbrace{\{b \in \Pi_{\mathcal{H}}(\mathcal{S}) : x \notin b \text{ and } b \cup \{x\} \in \Pi_{\mathcal{H}}(\mathcal{S})\}}_{\mathcal{B}'}|$
since $x \notin b$ for all $b \in \mathcal{B}'$, $\Pi_{\mathcal{B}'}(\mathcal{S} - \{x\}) = \mathcal{B}'$ $(*)$

$\ulcorner$ WHY $(*)$?
$\Pi_{\mathcal{B}'}(\mathcal{S} - \{x\}) = \{(\mathcal{S} - \{x\}) \cap b : b \in \mathcal{B}'\}$
where $\mathcal{B}' \stackrel{\text{def}}{=} \{b \in \Pi_{\mathcal{H}}(\mathcal{S}) : x \notin b, b \cup \{x\} \in \Pi_{\mathcal{H}}(\mathcal{S})\}$
So $b = h \cap \mathcal{S}$ for some $h \in \mathcal{H}$,
$x \notin b$, so $b \subseteq \mathcal{S} - \{x\}$
So $(\mathcal{S} - \{x\}) \cap b = b$ for all $b \in \mathcal{B}'$.
So $\Pi_{\mathcal{B}'}(\mathcal{S} - \{x\}) = \{\underbrace{(\mathcal{S} - \{x\}) \cap b}_{=b} : b \in \mathcal{B}'\} = \mathcal{B}'$ $\lrcorner$

$$\begin{aligned}\text{Now } |\Pi_{\mathcal{H}}(\mathcal{S})| &= |\Pi_{\mathcal{H}}(\mathcal{S} - \{x\})| + \underset{\text{\# collisions}}{|\mathcal{B}'|} \\ &\underset{\text{projections}}{=} |\Pi_{\mathcal{H}}(\mathcal{S} - \{x\})| + |\Pi_{\mathcal{B}'}(\mathcal{S} - \{x\})|\end{aligned}$$

It's enough to show $\mathcal{B}'$ has VC-dimension $d-1$, since then $|\Pi_{\mathcal{B}'}(\mathcal{S} - \{x\})| \leq \Phi_{d-1}(m-1)$ by IH.
Suppose $\mathcal{S}' \subseteq \mathcal{S} - \{x\}$ is shattered by $\mathcal{B}'$, then by def. of $\mathcal{B}'$, $\mathcal{S}' \cup \{x\}$ is shattered by $\mathcal{H}$ since any labeling of $\mathcal{S}'$ can be extended to two different labelings of $x$ by $\mathcal{H}$.
Therefore, $|\mathcal{S}' \cup \{x\}| \leq$ VC-dimension of $\mathcal{H}$, i.e. $d$.
Therefore, $|\mathcal{S}'| \leq d - 1$.
Since $\mathcal{S}'$ was an arbitrary set shattered by $\mathcal{B}'$, the VC-dim of $\mathcal{B}' \leq d - 1$. ■

# 4 Proof of sample complexity upper bound, part 1

Let $\mathcal{S}_1 = [(x^{(1)}, b^{(1)}), ..., (x^{(m)}, b^{(m)})]$ be examples drawn from $\mathcal{D}$, labeled by $c$.
Let $A$ be the event that $\mathcal{S}_1$ isn't an $\epsilon$-net for $\Delta_{\mathcal{H}}(c)$.
equiv. $A =$ "$\exists r \in \Delta_{\mathcal{H},\epsilon}(c)$ s.t. $\forall i, x^{(i)} \notin r$".

If we draw another independent set of $m$ labeled examples $\mathcal{S}_2$, for any fixed $r$ (witnessing $\mathcal{S}_1$ is not an $\epsilon$-net), in expectation $\epsilon m$ examples from $\mathcal{S}_2$ hit $r$.
By Chernoff bound, at least $\frac{\epsilon m}{2}$ examples from $\mathcal{S}_2$ hit $r$ w.p. $\geq 1 - e^{-\frac{\epsilon m}{8}}$.
So, as long as $m > \frac{8}{\epsilon} \ln 2$, this occurs w.p. $\geq \frac{1}{2}$.

So, the event
$B \stackrel{\text{def}}{=}$ "$\exists r \in \Delta_{\mathcal{H},\epsilon}(c) : [\forall x \in \mathcal{S}_1, x \notin r]$ and $\mathcal{S}_2$ hits $r \geq \frac{\epsilon m}{2}$ times"
has $Pr[B|A] \geq \frac{1}{2}$.

$B \subseteq A$, so $Pr[A \text{ and } B] = Pr[B]$
By def. of conditional prob.,

$\underbrace{Pr[B|A]}_{\geq \frac{1}{2}} \cdot Pr[A] = Pr[A \text{ and } B] = Pr[B]$

So $Pr[B] \geq \frac{1}{2} Pr[A]$
If we can show $Pr[B] \leq \frac{\delta}{2} \Rightarrow Pr[A] \leq \delta$
So, it is enough for us to show $Pr[B]$ is small.

So, e.g. given sets $\mathcal{S}_1$ & $\mathcal{S}_2$,

$$B|\mathcal{S}_1, \mathcal{S}_2 = \bigvee_{r \in \Pi_{\Delta_{\mathcal{H},\epsilon}(c)}(\underbrace{\mathcal{S}_1, \mathcal{S}_2}_{\substack{2m \\ \text{examples}}})} \left[ (\mathcal{S}_1 \cap r = \emptyset) \text{ and } |\mathcal{S}_2 \cap r| \geq \frac{\epsilon m}{2} \right]$$

By union bound,

$$Pr[B] \leq \underbrace{\mathbb{E}}_{\substack{(\mathcal{S}_1, \mathcal{S}_2) \\ \text{2m ex's} \\ \text{from D}}} \left[ \sum_{r \in \Pi_{\Delta_{\mathcal{H},\epsilon}(c)}(\mathcal{S}_1, \mathcal{S}_2)} Pr\left[ (\mathcal{S}_1 \cap r) = \emptyset \text{ and } |\mathcal{S}_2 \cap r| \geq \frac{\epsilon m}{2} \mid \mathcal{S}_1, \mathcal{S}_2 \right] \right]$$

**Lemma:** $\Pi_{\Delta_{\mathcal{H},\epsilon}(c)}(2m) \leq \Pi_{\Delta_{\mathcal{H}}(c)}(2m) = \Pi_{\mathcal{H}}(2m)$
**Pf of Lemma:**
Recall $\Delta_{\mathcal{H},\epsilon}(c) \subseteq \Delta_{\mathcal{H}}(c) \stackrel{\text{def}}{=} \{[h \neq c](x) : h \in \mathcal{H}\}$
Viewing these now as <u>vectors</u> over {0,1} indexed by $\mathcal{X}$ (i.e. functions on $\mathcal{X}$),
note that $[h \neq c] = \underbrace{h \oplus c}_{\substack{\text{pointwise} \\ \text{exclusive-or}}}$ ,
i.e. as vectors $\mathbf{h} + \mathbf{c}$ mod 2.

Therefore, for any two error regions $\mathbf{r}_1, \mathbf{r}_2 \in \Delta_{\mathcal{H}}(c)$,
for $\mathbf{h}_1, \mathbf{h}_2$, respectively,
$\mathbf{h}_1 = \mathbf{h}_2 \Leftrightarrow \mathbf{h}_1 + \mathbf{c} = \mathbf{h}_2 + \mathbf{c}$ mod 2
$\Leftrightarrow \mathbf{r}_1 = \mathbf{r}_2$
this gives a 1-1 correspondence between $\Delta_{\mathcal{H}}(c)$ & $\mathcal{H}$
So, $|\Pi_{\Delta_{\mathcal{H}}(c)}(\mathcal{S})| = |\Pi_{\mathcal{H}}(\mathcal{S})|$
So, $\Pi_{\Delta_{\mathcal{H}}(c)}(m) = \max_{\substack{\mathcal{S} \subseteq \mathcal{X} \\ |\mathcal{S}|=m}} |\Pi_{\Delta_{\mathcal{H}}(c)}(\mathcal{S})| = \max_{\substack{S \subseteq \mathcal{X} \\ |\mathcal{S}|=m}} |\Pi_{\mathcal{H}}(\mathcal{S})| = \Pi_{\mathcal{H}}(m)$

Furthermore, $\forall \mathcal{S}$, $|\Pi_{\Delta_{\mathcal{H},\epsilon}(c)}(\mathcal{S})| \leq |\Pi_{\Delta_{\mathcal{H}}(c)}(\mathcal{S})|$
So, $\Pi_{\Delta_{\mathcal{H},\epsilon}(c)}(m) \leq \Pi_{\Delta_{\mathcal{H}}(c)}(m)$ ■