CSE 513T Theory of Artificial Intelligence & Machine Learning February 15, 2018

Lecture 10

Lecturer: Brendan Juba *Scribe: Andrew Cukierwar*

Contents

1 Definitions and Notations, again **1**

2 Finish VC-dimension sample complexity bound **1**

3 Learning halfspaces **3**

4 Start learning vs. cryptography **3**

1 Definitions and Notations, again

Dichotomies For any class of representations, $\mathcal{H}$ over $\mathcal{X}, \mathcal{S} \subseteq \mathcal{X}$, the dichotomies of S realized by $\mathcal{H}$ are $\Pi_{\mathcal{H}}(\mathcal{S}) \stackrel{\text{def}}{=} \{(h(x_1), ..., h(x_m)) : h \in \mathcal{H}, \mathcal{S} = \{x_1, ..., x_m\}\}$ or identifying $h \in \mathcal{H}$ with the set $\{x \in \mathcal{X} : h(x) = 1\}$, $\Pi_{\mathcal{H}}(\mathcal{S}) = \{h \cap \mathcal{S} : h \in \mathcal{H}\}$.

Shattering If $\Pi_{\mathcal{H}}(\mathcal{S})$ is all subsets of $\mathcal{S}$, we say $\mathcal{S}$ is shattered by $\mathcal{H}$.

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

Error Regions For any target property c, the class of error regions of h wrt c is

$\Delta_{\mathcal{H}}(c) \stackrel{\text{def}}{=} \{[h \neq c](x) : h \in \mathcal{H}\}$

The error regions of probability $\geq \epsilon$ under a distribution $\mathcal{D}$ is denoted $\Delta_{\mathcal{H},\epsilon}(c)$.

ϵ-net For any $\epsilon > 0$, we say $\mathcal{S} \subseteq \mathcal{X}$ is an ϵ-net for $\Delta_{\mathcal{H}}(c)$ w.r.t $\mathcal{D}$ if for every $r \in \Delta_{\mathcal{H},\epsilon}(c)$ there is some $x \in r \cap \mathcal{S}$.

Growth Function The growth function of $\mathcal{H}$ is $\Pi_{\mathcal{H}}(m) \stackrel{\text{def}}{=} \max_{\substack{\mathcal{S} \subseteq \mathcal{X} \\ |\mathcal{S}|=m}} |\Pi_{\mathcal{H}}(s)|$.

2 Finish VC-dimension sample complexity bound

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

From last time, we proved:

Lemma 1: If $\mathcal{S}_1$ and $\mathcal{S}_2$ are sets of m examples drawn independently from $\mathcal{D}$ and labeled by c, the events $A =$ "$\mathcal{S}_1$ is not an ϵ-net for $\Delta_{\mathcal{H}}(c)$" and $B =$ "$\exists r \in \Delta_{\mathcal{H},\epsilon}(c) : [\mathcal{S}_1 \cap r = \emptyset]$ and $[|\mathcal{S}_2 \cap r| \geq \frac{\epsilon m}{2}]$" satisfy $Pr[A] \leq 2Pr[B]$.

Note if A does not occur, we can output any solution to $\text{Consis}_{\mathcal{H}}(\mathcal{S}_1)$. So, it is enough to show $Pr[B] \leq \frac{\delta}{2}$

Lemma 2: $\Pi_{\Delta_{\mathcal{H},\epsilon}(c)}(2m) \leq \Pi_{\Delta_{\mathcal{H}}(c)}(2m) = \Pi_{\mathcal{H}}(2m)$

Sauer's Lemma: If $\mathcal{H}$ has VC-dimension d, then $\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$, which is $\leq (\frac{\epsilon m}{d})^d$.

Suppose we have drawn $\mathcal{S} = (\mathcal{S}_1, \mathcal{S}_2)$, we can write $B|\mathcal{S}_1, \mathcal{S}_2$ as
$B|\mathcal{S}_1, \mathcal{S}_2 = \bigvee_{r \in \Pi_{\Delta_{\mathcal{H},\epsilon}(c)}(\mathcal{S}_1,\mathcal{S}_2)} [\mathcal{S}_1 \cap r = \emptyset] \wedge [|\mathcal{S}_2 \cap r| \geq \frac{\epsilon m}{2}]$
$Pr[B] = \mathbb{E}_{\mathcal{S}}[\mathbb{E}[B|\mathcal{S}]] \leq \mathbb{E}_{\mathcal{S}}[\sum_{r \in \Pi_{\Delta_{\mathcal{H},\epsilon}(c)}(\mathcal{S}_1,\mathcal{S}_2)} Pr[[\mathcal{S}_1 \cap r = \emptyset] \wedge [|\mathcal{S}_2 \cap r| \geq \frac{\epsilon m}{2}]|\mathcal{S}]]$
$\leq \mathbb{E}_{\mathcal{S}}|\Pi_{\Delta_{\mathcal{H},\epsilon}(c)}(\mathcal{S})| * 2^{\frac{-\epsilon m}{2}} \leq \Pi_{\mathcal{H}}(2m) * 2^{\frac{-\epsilon m}{2}} \leq (\frac{\epsilon m}{d})^d * 2^{\frac{-\epsilon m}{2}}$

This follows since each example in $\mathcal{S}_1$ and $\mathcal{S}_2$ is an independent draw from the same distribution $\mathcal{D}$. So, we could first draw $\mathcal{S}$ of size $2m$, and choose m of them uniformly at random to be $\mathcal{S}_1$ (the rest are $\mathcal{S}_2$). E.g. we randomly reorder $\mathcal{S}$, and call the first m elements $\mathcal{S}_1$, the second m elements $\mathcal{S}_2$.

Lemma 3: $\underset{\substack{\mathcal{S}_1,\mathcal{S}_2 \\ (\mathcal{S}_1,\mathcal{S}_2)=\mathcal{S}}}{Pr} [[\mathcal{S}_1 \cap r = \emptyset] \wedge [|\mathcal{S}_2 \cap r| \geq \frac{\epsilon m}{2}]|\mathcal{S}] \leq 2^{\frac{-\epsilon m}{2}}$

Proof: Note that $\mathcal{S} \cap r$ must be between m and $\frac{\epsilon m}{2}$, or else, respectively, the first or second condition cannot hold. The probability above is at most $\overline{Pr}[\mathcal{S}_1 \cap r = \emptyset \mid |\mathcal{S} \cap r| \geq \frac{\epsilon m}{2}]$
which is (let $l = \mathcal{S} \cap r$)...
$\frac{\text{Number of ways all } l \text{ elements of } \mathcal{S} \cap r \text{ end up in } \mathcal{S}_2}{\text{Number of ways all } l \text{ elements of } \mathcal{S} \cap r \text{ end up in } \mathcal{S}_1, \mathcal{S}_2} = \frac{\binom{m}{l}}{\binom{2m}{l}} = \frac{m!}{l!(m-l)!} \frac{l!(2m-l)!}{(2m)!} = \prod_{i=0}^{l-1} \frac{1}{2} = 2^{-l}$
Since $l \geq \frac{\epsilon m}{2}$, the lemma follows.
Q.E.D.

So, we only need to show $(\frac{\epsilon m}{d})^d * 2^{\frac{\epsilon m}{2}} \leq \frac{\delta}{2} \iff d \log(\frac{\epsilon m}{d}) - \frac{\epsilon m}{2} \leq \log \frac{\delta}{2}$.
Theorem follows from Lemma 4.

Lemma 4: If $m \geq \frac{12}{\epsilon}(d \log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta}))$, then $\frac{\epsilon m}{2} \geq d \log \frac{\epsilon m}{d} + \log \frac{2}{\delta}$

Proof: Since $1 + x \leq e^x$, for any $a, b > 0$, taking $x = ab - 1$ gives $\ln a + \ln b \leq ab - 1$.
Suppose we put $b = m$ and $a = \frac{\epsilon \ln 2}{4d}$
$\ln \frac{\epsilon \ln 2}{4d} + \ln m \leq \frac{\epsilon m}{4} * \frac{\ln 2}{d} - \ln e$
$[d \log_2 \frac{\epsilon \ln 2}{4d} + d \log_2 me] \leq \frac{\epsilon m}{4}$
So, it is enough to take $\frac{\epsilon m}{4} \geq d \log_2 \frac{1}{d} + \log \frac{2}{\delta} + d \log_2 \frac{4d}{\epsilon \ln 2}$
Since then $\frac{\epsilon m}{4} + \frac{\epsilon m}{4} \geq d \log_2 \frac{\epsilon \ln 2}{4d} + d \log_2 me + d \log_2 \frac{1}{d} + \log \frac{2}{\delta} + d \log_2 \frac{4d}{\epsilon \ln 2}$
$\frac{\epsilon m}{2} \geq d \log_2 \frac{me}{d} + \log \frac{2}{\delta}$
So it suffices to have $m \geq \frac{4}{\epsilon}[d \log \frac{1}{d} + \log \frac{2}{\delta} + d \log \frac{4d}{\epsilon \ln 2}]$ where $\log \frac{2}{\delta} \geq \log \frac{1}{\delta} + 1$ and $d \log \frac{1}{\epsilon} + d \log \frac{4}{\ln 2} \leq 2d$
So...if
$m \geq 4[3d \log(\frac{1}{\epsilon}) + 3 \log(\frac{1}{\delta})]$ where $3d \log(\frac{1}{\epsilon}) \geq d \log(\frac{1}{\epsilon}) + 2d$
Q.E.D.

Note:
1. Hanneke showed $m \geq \frac{1}{\epsilon}(d + \log \frac{1}{\delta})$ examples suffice. So the number of examples we need to learn class $\mathcal{H}$ of VC-dimension d is $O(\frac{1}{\epsilon}(d + \log \frac{1}{\delta}))$ (HW5...)
2. Although it can be easier computationally to learn a more expressive $\mathcal{H}$, it costs more data.

3 Learning halfspaces

HW4: Linear halfspaces in $\mathbb{R}^n$ have VC-dimension $n+1$.
Linear halfspaces functions $[\sum_{i=1}^{n} w_i x_i \geq t]$ (boolean functions) for $w_1, ..., w_n \in \mathbb{R}, t \in \mathbb{R}$
So, by our theorem: it suffices to find a halfspace consistent with $\frac{12}{\epsilon}((n+1)\log\frac{1}{\epsilon} + \log\frac{1}{\delta})$ examples to learn a halfspace.
Find $w_1, ..., w_n, t$ in $\mathbb{R}$ such that:
$w_1 x_1^{(i)} + ... + w_n x_n^{(i)} + (-1)t \geq 0$ for $b^{(i)} = 1$
$-w_1 x_1^{(i)} + ... + -w_n x_n^{(i)} + t \geq 0$ for $b^{(i)} = 0$
This is a homogeneous linear program, meaning of the form $A(\overrightarrow{w}, t) \leq 0$....can solve in polynomial time.
Notice...given any homogeneous linear program $A(\overrightarrow{w}, t) \leq 0$, by taking the ith example to be the ith row of A ("A_i"), $b^{(i)} = 0$ and $odd(\overrightarrow{0}, 1)$, any solution to $\text{Consis}_{\text{Halfspace}}$ on this set must satisfy
(1) $\overrightarrow{0}\overrightarrow{w} - t \geq 0 \Rightarrow t \leq 0$
(2) for every ith row: $A_i\overrightarrow{w} + t \geq 0$ and $A_i\overrightarrow{w} - t \leq 0 \Rightarrow A_i\overrightarrow{w} \leq 0$
So: Proper PAC-learning of halfspaces is equivalent to solving homogeneous linear programs.
Example $\leftrightarrow$ Constraint in linear program

4 Start learning vs. cryptography

What we will see next:
Classes like Booelan Circuits cannot be learned by polynomial-time algorithms.

Their NP-Completeness is not known! Instead we will show learning would break cryptographic assumptions. Learning would allow us to solve some problems used by cryptography - such as factoring, shortest lattice vector, discrete, roots, etc. - and break cryptographic systems.