CSE 513T Theory of Artificial Intelligence & Machine Learning February 19, 2015

Lecture 12

Lecturer: Brendan Juba *Scribe: Kai Zhang*

1 Today's topic

- Finish VC-Dimension sample complexity bound.
- start representation-independent hardness crypographic hardness vs. NP-hardness.

2 VC-Dimension sample complexity bound

Theorem *let c be any Boolean concept,let H be any repersentation class of VC-dimension d. Then any solotion h to $Consis_H$ on a data set of $m \geq C_\varepsilon(d\log\frac{1}{\varepsilon}+\log\frac{1}{\delta})$ examples drawn independence from D and labeled by C satisfies $Pr_{x\in D}[h(x)=c(x)] \geq 1-\varepsilon$,w.p.$1-\delta$*

Definition *The class of error regions of H relative to c, $\triangle_H(c)=\{[h(x)\neq c(x)];h\in H\}$,$\triangle_{H,\varepsilon}(c)=\{r\in\triangle_H(c)=Pr_{x\in D}[x\leq r]\geq\varepsilon\}$,$S\subset X$,is an ε-net for $\triangle_H(c)$ wrt D if $\forall r\in\triangle_{H,\varepsilon}(c)$, $\exists x\in r\bigcap s$*
The growth fuction of H is $\pi_{H(m)=\max_{s,|s|=m}|\pi_H(s)|}$

Proof proof of VC bound:
From the last time H suffice to show that
$Pr_{S_1,S_2}[\exists r\in\triangle_{H,\varepsilon}(c):[\forall_{x,b}\in S,x\notin r]and[|S_2\|\bigcap r]\geq\frac{\varepsilon m}{2}times]<\frac{\delta}{2}$
($\rightarrow$ wp.$>1-\delta$,S_1 is an ε-net for H $\rightarrow$ any solution to $Consis_H(S_1)$satisfies $Pr[h=c]\geq 1-\varepsilon$)
"Laws of total probability",break up in terms of $S(=S_1\bigcup S_2)$
$E_{S,|s|=2m}[Pr_{S_1S_2}[\exists r\in\triangle_{H,\varepsilon}(c):[S_1\bigcap r=\phi]and[|S_2\bigcap r|\geq\frac{\varepsilon m}{2}|S=S_1\bigcup S_2]]$
$\leq E_{S,|s|=2m}[\sum_{r\in\triangle_{H,\varepsilon}(c)}Pr_{S_1S_2}[\exists r\in\triangle_{H,\varepsilon}(c):[S_1\bigcap r=\phi]and[|S_2\bigcap r|\geq\frac{\varepsilon m}{2}|S=S_1\bigcup S_2]]$ ■

Lemma $E_{S,|s|=2m}[Pr_{S_1S_2}[\exists r\in\triangle_{H,\varepsilon}(c):[S_1\bigcap r=\phi]and[|S_2\bigcap r|\geq\frac{\varepsilon m}{2}|S=S_1\bigcup S_2]]\leq 2^{-\frac{\varepsilon m}{2}}$
proof:
consider the following experiment.Suppose I sample S_1+S_2 in the following way, I first draw a set S of 2m examples independently from D,then randomly choose since S consisted of 2m independent draws from D,S_1 is just an independent random choice of m out of these 2m, S_1+S_2 have the same distribution as before(eg,conditioning on the values of S_1 doesn't change the distribution of S_2).
In order for the event to have nonzero probability,$|S\bigcap r|\geq\frac{\varepsilon m}{2}$(or else second condition never satisfied). so the probability is at most $Pr_{S_1}[S_1\bigcap r=\phi|S_1\subset S,|S\bigcap r|\geq\frac{\varepsilon m}{2}]$.Let $l=|S\bigcap r|$,equivalently,let's think of the first m positions as S_1,and choose the l positions (out of 2m)uniformly at random.
$Pr_{S_1}[S_1\bigcap r=\phi|S_1\subset S,|S\bigcap r|\geq\frac{\varepsilon m}{2}]\leq\frac{\binom{m}{l}}{\binom{2m}{l}}=\prod_{s=0}^{l-1}\frac{m-1}{2m-1}\leq\frac{1}{2^l}$,*where* $l\geq\frac{\varepsilon m}{2}$

Lemma $\Pi_{\triangle_H(c)}(2m)\leq\Pi_H(2m)$
proof:
recall that $\triangle_H(c)$ consists of the sets/functions $[h(x)\neq c(x)]$ (here,restricted to a set S of 2m points).Thinking of the labelings of S as vectors indexed by the elements of S,$[h(x)\neq c(x)]=h(x)+c(x) \pmod 2$, so for any labelings of S by $[h(x)\neq c(x)]$
There is a distinct labeling of S by h since for any two $h,h'\in H$,if $[h(x)\neq c(x)]=[h'(x)\neq c(x)]$,then$h(x)+c(x)=h'(x)+c(x) \pmod 2\rightarrow h(x)\neq h'(x)$. And conversely,if $h(x)=h'(x)$,obtain$[h(x)\neq c(x)]$by adding c, $[h(x)\neq c(x)]=[h'(x)\neq c(x)]$.So the number of labelings is all same for S,and so $\Pi_{\triangle_H(c)}(2m)\leq\Pi_H(2m)$

since H has VC-dimension d.$m > d$,Sauer's Lemma get that:
$|\Pi_{\triangle_H(c)} = (S)| \leq |\Pi_H(S)| \leq (\frac{2\varepsilon m}{d})^d$ for all $|S| = 2m$, So overall,$Pr[...\text{"}B\text{"}...] \leq 2^{-\frac{\varepsilon m}{2}}(\frac{2me}{d})^d$,
$\frac{2}{\delta}(\frac{2me}{d})^d \leq 2^{\frac{\varepsilon m}{2}}$,
$\log\frac{2}{\delta} + d\log\frac{2me}{d} \leq \frac{\varepsilon m}{2}$
So we just need to find the paint where these are equal.So look at $m > \frac{2}{\varepsilon}[d + \log\frac{2}{\delta} - \log\frac{d}{e}]$
claim $m > \frac{16}{\varepsilon}(d\log\frac{16\varepsilon}{\varepsilon} + log\frac{2}{\delta})$is sufficient
check plugging in this m into
$\frac{2}{\varepsilon}(log\frac{2}{\delta} + d\log\frac{2me}{d})$
$= \frac{2}{\varepsilon}(\log\frac{2}{\delta} + d\log(\frac{2e}{d}\frac{16}{\delta}(d\log\frac{16e}{\varepsilon} + \log\frac{2}{\delta})))$
$\leq d\log(\frac{2e}{d}\frac{16}{\varepsilon}d\log\frac{16e}{\varepsilon} + \log\frac{2}{\delta})$
$\leq \frac{16e}{\varepsilon}(\log\frac{2}{\delta} + d\log\frac{16e}{\varepsilon}) = m$
Application Learning linear halfspaces
Fact VC-dimension of halfspaces in $\mathbb{R}^n$ is $n+1$
$m \sim \frac{1}{\varepsilon}(n\log\frac{1}{\varepsilon} + \log\frac{1}{\delta})$ exmaples suffice to fit a halfspace in $\mathbb{R}^n$
the consistency problem for halfspaces
Find $w_1, w_2, ..w_n, t \in \mathbb{R}$ s.t. linear program feasiblity
$\{w_1x^{(1)} + ..w_nx^{(n)} \geq tforeachx(t)s.t.b(t) = 1$
$w_1x^{(1)} + ..w_nx^{(n)} \geq tforeachx^{(t)}s.t.b^{(t)} = 0\}$
LP feasiblity
find $X \in \mathbb{R}^n$,s.t.$Ax \leq bwhereA \in \mathbb{R}^{m*n}, b \in \mathbb{R}$

Theorem *(Kachiyan/Karmarkar,etc)There're poly-time1 algorithms for Lp feasiblity.*
Remenber proper learning algorithms can be used to solve consistency, for a suitable choice of data set,learning halfspace is"essientially " the general case of LP feasiblity.

3 Representation-independent hardness crypographic hardness vs.NP-hardness

We saw previously that proper learning of 3-term DNF is NP-complete,but 3-term DNF can be efficiently improperly learned by 3-CNFs.
Are there concept classes that can not even be learned improperly?
First instinct show that such a problem is NP-complete.But,we don't know how to show anything like this!
Instead,we will show that such algorithms can be used to violate the kinds of assumptions underlying crypography.
Crypotography is based on the hardness of specific problems ,which are believed not to be NP-complete.
Basic tension between crypotography and learning

Theorem $Meta - theorem(Goldreich - Goldwasser - Micali'86, Hgstad - Impagliazzo - Levin - Luby'99)$
If crypotography is possible at all, then there is no PAC learning algorithms for Boolean Circuts.

Theorem *Partial Converse(Blum-Furst-Kearns-Lipton'94,Regev'09)*
Hard learning problems can often be used to design crypotographs.

Intuition Suifficiently hard learning problem$(x^{(1)}, c(x^{(1)}), (x^{(2)}, c(x^{(2)})...(x^{(m)}, c(x^{(m)})$ is distingguishable from $(x^{(1)}, b^{(1)}), (x^{(2)}, b^{(2)})...(x^{(m)}, b^{(1m)})$where each $b^{(i)}$ is an independent,random bit.
Someone who knosw the secret c can distinguish these.(issues are sampling good choices of c,good D,etc.)

4 Next time

Specific assumption that would be violated by efficient PAC-learning of Boolean Circuits.