CSE 513T Theory of Artificial Intelligence & Machine Learning Feburary 17th, 2015

Lecture 11

Lecturer: Brendan Juba *Scribe: Yao Zhou*

1 Today's Topics

1. Proof of Sauer's Lemma
2. VC-dimension sample complexity upper bound

2 Warm-up Problem

1. $\binom{k}{l} + \binom{k}{l-1} = ?$
$\binom{k}{l}$: don't choose the element
$\binom{k}{l-1}$: do choose the element
We can recover the choice based on whether or not "k+1" in set. So the term count disjunctions.
Thus:$\binom{k}{l} + \binom{k}{l-1} = \binom{k+1}{l}$

2. $\Sigma_{i=0}^{k}\binom{k}{i} = ?$
•Solution 1:
choose subsets of all,size=2^k
$\binom{k}{0} + \binom{k}{1} + ... + \binom{k}{k-1} + \binom{k}{k} = 2^k$

•Solution 2:
We know:$(1+x)^k = \Sigma_{i=0}^{k}\binom{k}{i}x^i$
Here $x = 1$ and we get:
$\Sigma_{i=0}^{k}\binom{k}{i} = 2^k$

3 Definition from previous lecture

For concept class C over X, $S \subseteq X$, the dichotonics of S realized by C

$$\Pi_C(S) = \left\{[c(x^{(1)}), c(x^{(2)}), ..., c(x^{(n)})]; c \subseteq C\right\}$$

or identifying labeling $\left\{x^{(1)}, x^{(2)}, ..., x^{(n)}\right\}$ sets, $\Pi_C(S) = \{c \bigcap S, c \subseteq C\}$. If $\Pi_C(S)$ contains all labeling of S, we say S is shattered by C. The VC-dimension of C is the size of the largest set shattered by C. The growth function for C:

$$\Pi_C(m) = max_{S,|S|=m}\,|\Pi_C(S)|$$

For any target concept c of the class of error regions of H with regards to c, denoted $\Delta_H(c)$(with regards to D) is $\{r \in \Delta_H(c) : Pr_{x\in D}[x \in r] \geq \varepsilon\}$.

For any $\varepsilon > 0$, a set $S \subseteq X$ is an ε-end for $\Delta_H(c)$ with regards to D if for every $r \in \Delta_H(c)$ there is some $x \in r \bigcap S$.

4 Sauer's Lemmas

If VC-dimension of H is d, then for every m,

$$\Pi_H(m) \leq \Sigma_{i=0}^{d} \binom{m}{i}$$

($O(m^d)$ for $m > d$; for $m \leq d, \Pi_H(m) = 2^m$)

4.1 Lemma 1

For any $m, d \in N$, define $\Phi_d(m)$ inductively by: $\Phi_0(m) = \Phi_d(0) = 1$, otherwise $\Phi_d(m) = \Phi_{d-1}(m-1) = \Phi_d(m-1)$. If VCD($H$)=d, then: $\Pi_H(m) \leq \Phi_d(m)$.

Proof By induction on m,d

Base: $d = 0$ or $m = 0$. If $d = 0$,H must only contain constant function $\Pi_H(m) = 1 \leq \Phi_0(m)$; if $m = 0$, the only subset is ϕ so

$$\Pi_H(0) = |\Pi_H(\phi)| = 1 \leq \Phi_d(0)$$

Induction hepothesis: for any m', d', such that $m' < m$ or $d' < d$, $\Pi_H(m') \leq \Phi_{d'}(m')$

Induction step: Let any set S of size m be given, H of VC-dimension d. Choose any $x \in S$, since $|S - \{x\}|$ has size $m-1$, by induction hypothesis $|\Pi_H(S - \{x\})| \leq \Phi_d(m-1)$, $\Pi_H(S - \{x\})$(set of all labelings of $S - \{x\}$) contains all labelings of S, projected down to $S - \{x\}$.
So

$$\Pi_H(S) \leq |\Pi_H(S - \{x\})| + |\Pi_B(S - \{x\})|$$

where $B = \{r \in \Pi_H(S) : x \in r, r - \{x\} \in \Pi_H(S)\}$.

Consider $\Pi_B(S - \{x\})$, VCD(B)$\leq d-1$.
why?: Suppose $S' \subseteq S - \{x\}$ is shattered by B. Then $S' \bigcup \{x\}$ is shattered by H, since both $r \in S', r \bigcup \{x\} \in \Pi_H(S)$. In other words, x can take two labels.

So: VCD(B)$\leq$VCD(H)-1.
Since for every arbitary S, $2\,|\Pi_B(S - \{x\})| \leq |\Pi_H(S)|$
$\Rightarrow d' = d - 1$
So $|\Pi_B(S - \{x\})| \leq \Phi_{d-1}(m-1)$ by induction hypothesis, and hence:

$$\Pi_H(S) \leq \Phi_d(m-1) + \Phi_{d-1}(m-1) = \Phi_d(m) \text{ (by definition)}$$

4.2 Lemma 2

$$\Phi_d(m) = \Sigma_{i=0}^{d} \binom{m}{i} \leq \left(\frac{me}{d}\right)^d \quad \text{for } m > d$$

Proof by induction on m and d

Base: $m = 0, \Sigma_{i=0}^{d} \binom{0}{i} = l = \Phi_d(0)$, $d = 0, \Sigma_{i=0}^{0} \binom{m}{i} = \binom{m}{0} = 1 = \Phi_0(m)$.

Induction hypothesis: As in claim.

Induction step:

$$\begin{aligned}\Phi_d(m) &= \Phi_d(m-1) + \Phi_{d-1}(m-1) \\ &= \Sigma_{i=0}^{d}\binom{m-1}{i} + \Sigma_{i=0}^{d-1}\binom{m-1}{i} \\ &= \Sigma_{i=0}^{d}\binom{m-1}{i} + \Sigma_{i=1}^{d}\binom{m-1}{i-1} \\ &= \Sigma_{i=0}^{d}(\binom{m-1}{i} + \binom{m-1}{i-1}) \\ &= \Sigma_{i=0}^{d}\binom{m}{i}\end{aligned} \tag{1}$$

Complete inductor:
For the second part, for $m > d$, $\frac{d}{m} < 1$

$$\begin{aligned}(\frac{d}{m})^d\Sigma_{i=0}^{d}\binom{m}{i} &\leq \Sigma_{i=0}^{d}(\frac{d}{m})^i\binom{m}{i} \\ &\leq \Sigma_{i=0}^{m}(\frac{d}{m})^i\binom{m}{i} \\ &= (1+\frac{d}{m})^m \leq e^d\end{aligned} \tag{2}$$

for $d > 0$

$$\Sigma_{i=0}^{d}\binom{m}{i} \leq (\frac{me}{d})^d$$

5 VC-dimension sample complexity bound

Theorem:

Let c be any boolean concept, H be any representation class of VC-dimension d. Then any solution to $Consis_H$ on a data set of $m \geq C_0(d\log\frac{1}{\varepsilon} + \log\frac{1}{\delta})$ examples (for $C_0 > 0$) drawn independently from some D and labeled by c satisfies $Pr_{x\in D}[h(x) = c(x)] \geq 1-\varepsilon$ with probability $1-\delta$(over data set).

Strategy:

It suffices to show that with probability $1-\delta$ the data set forms an ε-net for $\Delta_H(c)$. Since h has small VC-dimension, the number of dichotonics of H on a sample of size $m > d$ is bounded by $\Pi_H(m) = O(m^d)$. So we'll be able to afford an union bound over the labelings. H gives to a sample of size m.

Trick : velating the probelm : the data forms ε − net to its behavior on a small sample.

Proof

Let $S = [(x^{(1)}, b^{(1)}), (x^{(2)}, b^{(2)}), ..., (x^{(m)}, b^{(m)})]$ of examples drawn form D and labeled by c (a random variable). Let A be the event "$\exists r \in \Delta_{H,\varepsilon}(c)$ such that $\forall i, x^{(i)} \notin r$". S is not an ε-net for $\Delta_H(c)$.

We need to rewrite the condition in terms of a small sample.
Notice, if we draw another set of m examples from D, S_2, and if A holds, then for any fixed $r \in \Delta_{H,\varepsilon}(c)$ (that missed S_1) will be hit εm times by S_2 in expectation.(Since each draw hits r with probability $\geq \varepsilon$).

A Chernoff bound tells us that the probability that r is hit less than $\frac{\varepsilon m}{2}$ times by S_2 is bounded by

$$e^{-\varepsilon m(\frac{1}{2})^2/2} = e^{-\varepsilon m/8}$$

So as long as $m > \frac{\delta}{\varepsilon} \ln 2$(it will be even larger), this probability is $< \frac{1}{2}$. So the event

$$B = \text{“}\exists r \in \Delta_{H,\varepsilon}(c) : \left[\forall x \in S_1, x \notin r\right] \ \& \ \left[\left|r \bigcap S_2\right| \geq \frac{\varepsilon m}{2}\right]\text{”}$$

Notice since $A \subseteq B$, $Pr[A \ \& \ B] = Pr[B]$, And $Pr[B/A] \geq \frac{1}{2}$.

$$\frac{Pr[B \ \& \ A]}{Pr[A]} \longrightarrow Pr[B] \geq \frac{Pr[A]}{2}$$

It now suffices to show $Pr[B] \leq \frac{\delta}{2}$
Now:

$$B|S_1, S_2 = \bigvee_{r \in \Pi_{\Delta_{H,\varepsilon}(c)}(S_1 \bigcap S_2)} \left(\left[S_1 \bigcap r = \phi\right] \ and \ \left[\left|S_2 \bigcap r\right| > \frac{\varepsilon m}{2}\right]\right)$$

$$Pr[B] \leq E_{S,|S|=2m}\left[\Sigma_{r \in \Pi_{\Delta_{H,\varepsilon}(c)}(S)} Pr_{S_1,S_2}\left[\left[S_1 \bigcap r = \phi\right] \ and \ \left[\left|S_2 \bigcap r\right| > \frac{\varepsilon m}{2}\right] \middle| S_1 \bigcup S_2 = S\right]\right]$$

Lemma

$$Pr_{S_1,S_2}\left[\left[S_1 \bigcap r = \phi\right] \ and \ \left[\left|S_2 \bigcap r\right| > \frac{\varepsilon m}{2}\right] \middle| S_1 \bigcup S_2 = S\right] \leq 2^{-\varepsilon m/2}$$