CSE 513T Theory of Artificial Intelligence & Machine Learning October 6, 2016

Lecture 12

Lecturer: Brendan Juba *Scribe: Zongyi Li Edit: Josh Hansen*

1 Today's Topics

1. VC-dimension sample complexity upper bound
2. Application: halfspaces (linear threshold functions) are PAC-Learnable

2 Preliminary Definitions (from the last 2 lectures)

Dichotomy For any class of representations C over X and $S \subset X$, the dichotomies of S realized by C are:

$$\Pi_C(S) = \{(c(x^{(1)}), ..., c(x^{(m)})); c \subset C, S = \{x^{(1)}, ..., x^{(m)}\}\}$$

Shattered If $\Pi_C(S) = \{0,1\}^{|S|}$, then we say S is shattered by C.

VC-Dimension (VCD) The VC-Dimension of C is the size of the largest S shattered by C.

Error Region For any target concept c, the class of error region of H with respect to c is:

$$\triangle_H(c) = \{[h \neq c](x), h \in H\}$$

The error region of probability $\geq \epsilon$ under a fixed distribution D is:

$$\triangle_{H,\epsilon}(c) = \{r \in \triangle_H(c) : Pr_{x \in D}[x \in r] \geq \epsilon\}$$

Growth Function The growth function of a representation class H is:

$$\Pi_H(m) = max_{S \subset X, |S|=m} |\Pi_H(S)|$$

Sauer's Lemma If H has VC-dimension d, then:

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

No Free Lunch Theorem Learning C of VC-dimension d requires $\Omega(d/\epsilon)$ examples for $S < S_0$, where S_0 characterizes the "sample complexity."

3 VC-Dimension Sample Complexity Upper Bound Theorem

3.1 Claim

Let c be any property and H be any representation class of VC-dimension d. Then any solution h to $Consis_H$ on a data set consisting of examples drawn independently from some distribution D, and labeled by c will satisfy $Pr_{x \in D}[h(x) = c(x)] \geq 1 - \epsilon$ with probability $1 - \delta$ as long as $m \geq c_0 * 1/\epsilon * (d * log(1/\epsilon) + log(1/\delta))$ for some constant $c_0 > 0$ (Hanneke, 2016: actually $O(1/\epsilon(d + log(1/\delta)))$ suffices).

3.2 Strategy

Recall that it suffices to show that the data set forms an "ϵ-net" for H (i.e. for every $r \in \triangle_{H,\epsilon}(c), \exists x \in r$). Since H has a small VC-dimension, the number of functions that can appear on a finite sample is bounded by $O(m^d)$ by Sauer's Lemma. We can afford to take a union bound over those distinct labelings on the examples. The key trick is to relate the behavior of $h \in H$ on D to it's behavior on a small sample.

3.3 Proof

Let $S_1 = [(x^{(1)}, b^{(1)}), ..., (x^{(m)}, b^{(m)})]$ be our examples drawn from D and labeled by c. We can write the event $A =$ "S_1 is not an ϵ-net for $\triangle_H(c)$" = "$\exists r \in \triangle_{H,\epsilon}(c)$ such that $\forall i, x^{(i)} \neq r^{(i)}$."

We need to rewrite this condition in terms of another small "test" sample of another m examples. Notice, if we draw another set S_2 of m examples from D, then for any fixed error region ($r \in \triangle_{H,\epsilon}(c)$), at least ϵm of the samples have $c(x) \neq h(x)$.

A Chernoff bound then guarantees that the probability that fewer than $\epsilon m/2$ examples hit r is at most $e^{-\epsilon m/8}$. So, as long as $m > (8/\epsilon)ln2$, this occurs with probability $< 1/2$.

Let's consider the event $B =$ "$\exists r \in \triangle_{H,\epsilon}(c), \forall x^{(i)} \in S_1, x^{(i)} \notin r, S_2$ hit $r > \epsilon m/2$ times."

We have $Pr_{S_1,S_2}[B|A] > 1/2$, where $B \subseteq A$ so $Pr[B \cap A] = Pr[B]$. Note: $Pr[B|A] = Pr[B \cap A]/Pr[A]$.

Therefore, $Pr[B] > \frac{1}{2}Pr[A]$. It therefore suffices to show $Pr[B] < \delta/2$ since $Pr[A] < \delta$. Note: B is defined in terms of the behavior of $h \in H$ on the two samples S_1 and S_2.

Let's rewrite B, given a fixed S_1 and S_2:

$$B|S_1, S_2 = \bigcup([S_1 \cap r = \emptyset] \wedge [|S_2 \cap r| \geq \epsilon m/2]), r \in \Pi_{\triangle_{H,\epsilon}}(S_1, S_2)$$

$$Pr[B] = E_{S_1,S_2}[Pr[B|S_1, S_2]]$$

By a union bound over set of error regions on S_1 and S_2 ($\Pi_{\triangle_{H,\epsilon}}(S_1, S_2)$),

$$Pr[B] \leq E_{S_1,S_2}\big[\sum_r Pr[(S_1 \cap r) = \emptyset \wedge |S_2 \cap r| \geq \epsilon m/2 | S_1, S_2]\big]$$

First choose a combined list S of $2m$ examples and then randomly split S into S_1 and S_2, each with m examples. Note: since S_1 and S_2 consisted of m independent draws of examples from D labeled by c, this gives the same distribution on S_1 and S_2. So,

$$Pr[B] \leq E_{S \in D^{2m}}\big[\sum_r Pr_{S_1,S_2}[(S_1 \cap r) = \emptyset \wedge |S_2 \cap r| \geq \epsilon m/2]\big]$$

3.4 Lemma 1

$$Pr_{S_1,S_2}[(S_1 \cap r) = \emptyset \wedge |S_2 \cap r| \geq \epsilon m/2 | S = S_1 \cup S_2] \leq 2^{-\epsilon m/2}$$

3.5 Lemma 1 Proof

If we don't have $\epsilon m/2$ examples in r in S, then this event does not occur so it's probability is bounded by

$$Pr[S_1 \cap r = \emptyset || S \cap R| > \epsilon m/2]$$

We have $l = |S \cap R|$ points in our random split. All of l fall into S_2, which is size m. If we focus on the l locations where points of $S \cap r$ end up between the list S_1, S_2 in $\binom{m}{l}$ of these, all l end up in S_2 out of $\binom{2m}{l}$ possible locations. So, the probability is at most $\frac{\binom{m}{l}}{\binom{2m}{l}}$.

$$\frac{\binom{m}{l}}{\binom{2m}{l}} = \prod_{i=0}^{l-1} \frac{m-i}{2m-i} < \frac{i}{2} l \leq 2^{-\epsilon m/2}$$

3.6 Lemma 2

$$\Pi_{\triangle_{H,\epsilon}(c)}(2m) \leq \Pi_{\triangle_H(c)}(2m) = \Pi_H(2m)$$

3.7 Lemma 2 Proof

Recall that $\triangle_H(c) \geq \triangle_{H,\epsilon}(c$ consists of the sets $[h \neq c] : h \in H$. Thinking of these as vectors, indexed by x, we note that $[h \neq c] = \mathbf{h} + \mathbf{c} mod 2$. Therefore, for any 2 dichotomies r_1 and $r_2 \in \triangle_H(c)$, given by $[h_1 \neq c]$ and $[h_2 \neq c]$ for h_1 and $h_2 \in H$

$$\mathbf{h}_1 \neq \mathbf{h}_2 \Leftrightarrow \mathbf{h}_1 + c \neq \mathbf{h}_2 + c mod 2 \Leftrightarrow \mathbf{r}_1 \neq \mathbf{r}_2$$

So, for any S,

$$|\Pi_{\triangle_H(c)}(S)| = |\Pi_H(S)|$$

$$\Pi_{\triangle_H(c)}(2m) = max_{S:|S|=2m} |\Pi_{\triangle_H(C)}(S)| = \Pi_H(2m)$$

Also, since for every S, $\Pi_{\triangle_{H,\epsilon}(c)}(S) \subseteq \Pi_{\triangle_H(c)}(S)$,

$$\Pi_{\triangle_{H,\epsilon}(c)}(2m) = max_{S:|S|=2m} |\Pi_{\triangle_{H,\epsilon}(c)}(S)| \leq |\Pi_{\triangle_H(c)}(S)| = \Pi_{\triangle_H(c)}(2m)$$

3.8 Main Proof Continued

By Lemma 1,

$$Pr[B] \leq E_S[\sum_r 2^{-\epsilon m/2}] \leq 2^{-\epsilon m/2} E_S[\Pi_{\triangle_{H,\epsilon}(c)}(2m)] \leq 2^{-\epsilon m/2} E_S[Pi_{\triangle_H}(2m)]$$

By Sauer's Lemma, that value is $\leq 2^{-\epsilon m/2} \Pi_H(2m) \leq (\epsilon m/d)^d 2^{-\epsilon m/2}$.

We only need to show $\epsilon m/2 \geq d * log(\epsilon m/d) + log(2/\delta)$ for our m.

3.9 Claim

The following suffices:

$$m \geq \frac{4}{\epsilon ln2}(d * ln(\frac{e}{d}) + d * ln(\frac{4d}{\epsilon ln2}) - d + ln(\frac{2}{\delta}))$$

Recall that since $1 + x \leq e^x$, taking the natural log of each side and setting $x = ab - 1$, results in $ln(a) + ln(b) \leq ab - 1$ for $a, b > 0$.

Let's use $a = \frac{\epsilon ln2}{4d}$ and $b = m$. We find

$$ln(m) \leq \frac{\epsilon ln2}{4d} m + \frac{4d}{\epsilon ln2} - 1$$

$$\epsilon\frac{m}{2} \geq d * log(\frac{em}{d}) + log(\frac{2}{\delta})$$

Examining the bound on m,

$$m \geq \frac{4}{\epsilon}(d * log(\frac{4}{\epsilon ln2}) + log(\frac{2}{\delta}))$$

Taking $c_0 = 12$ completes the theorem.