Foundations of Data Science & Machine Learning

Summary — Week 06
Devansh Singh Rathore
111701011

B.Tech. in Computer Science & Engineering
Indian Institute of Technology Palakkad

April 14, 2021

Abstract

We discuss Sauer's Lemma and then use its results for our on going calculation. Then we discuss the case PAC learnability when VC dimension of hypothesis class is ∞. Finally, we try to reduce the sample complexity for finite VC dimension cases.

1 Sauer's Lemma

Lemma: If a hypothesis class H has a finite VC Dimension d, then

$$\boxed{g_H(n) \le \sum_{i=0}^{d} \binom{n}{i} = \binom{n}{\le d}}$$

Proof: (Induction on n)

Let g(d, n) = max $g_H(n)$: H has VC dimension d

We will show g(d, n) $\le \binom{n}{\le d}$

NOTE: When trying to prove for some n, we assume the claim for all d and n-1.

Base case: (n = 0)

if there are no points, the only possible hypothesis class is ϕ and hence

$$\forall d, g(d, 0) = 1 = \binom{0}{\le d}$$

(**Exercise:** prove for n=1 case: g(0,1) = 1, g(d,1) = 2 $\forall d > 0$)

Induction Hypothesis:

$$\forall d, g(d, n-1) \le \binom{n-1}{\le d}$$

Induction Step(n):

H : any hypothesis class with VC dim. d

X_n : any set of n points from X, $X_n \in \binom{X}{n}$

$H_n = H|_{X_n} = \{h|_{X_n} : h \in X_n\}$

Claim: $|H_n| \le \binom{n}{\le d}$

Let $X_n = \{x_1, x_2, ..., x_n\}$ and $X_{n-1} = \{x_1, x_2, ..., x_{n-1}\} = X_n \backslash \{x_n\}$

Two functions f, g $\in H_n$ are said to be equivalent if $f|_{X_{n-1}} = g|_{X_{n-1}}$ i.e. $f \approx g$ iff $f(x_i) = g(x_i) \forall i \in [n-1]$.

This is an equivalence relation and $\forall f \in H_n, |[f]| = 1 or 2$
Let $\alpha = \#$ 2-sized eq. classes and
$\beta = \#$ 1-sized eq. classes
So, $|H_n| = 2\alpha + \beta$, whereas $|H_{n-1}| = \alpha + \beta$
Since $H_{n-1} = H_n|_{X_{n-1}} = H|_{X_{n-1}}$,

$$|H_{n-1}| \leq g(d, n-1) \leq \binom{n-1}{\leq d}$$

$$i.e.\ \alpha + \beta \leq \binom{n-1}{\leq d}$$

Consider $H'_{n-1} \subseteq H_{n-1}$ obtained by restricting 2-sized eq. classes of H_n, hence $|H'_{n-1}| = \alpha$.

H'_{n-1} cannot shatter a subset $T \subseteq X_{n-1}$ of size larger than d-1, because then $T \cup \{x_n\}$ is shattered by H_n.

Hence VC dim.$(H'_{n-1}) \leq d - 1$

$$|H'_{n-1}| \leq g(d-1, n-1) \leq \binom{n-1}{\leq d-1}$$

$$i.e.\ \alpha \leq \binom{n-1}{\leq d-1}$$

$$|H_n| = 2\alpha + \beta$$
$$= (\alpha + \beta) + \alpha$$
$$\leq \binom{n-1}{\leq d} + \binom{n-1}{\leq d-1}$$
$$\leq \binom{n}{\leq d}$$

Hence Proved.

1.1 Converse of Sauer's Lemma

VC Dim(H) $= \infty \Rightarrow g_H(n) = ?$

VC Dim(H) $= \infty \Rightarrow \exists$ an infinite set $T \subseteq X$ shattered by H.
Let $T_n \in \binom{T}{n}$, then T_n is also shattered by H.
Hence $|H|_{T_n}| = 2^n$
So $g_H(n) \geq 2^n$
But # of binary functions on any n-set $\leq 2^n$
So $g_H(n) = 2^n$

$$Hence,\ VCDim(H) = \infty \Rightarrow g_H(n) = 2^n$$

Obs.: $g_H()$ is either polynomially bounded or otherwise exponential, nothing in b/w.

2 Finishing the Calculation

$$g_H(n) 2^{-\epsilon n/2} \leq \delta/2$$
$$\binom{2n}{d} 2^{-\epsilon n/2} \leq \delta/2$$
$$Upon\ solving\ gives,\ n \geq 4/\epsilon(2d\, log(1/\epsilon) + log(2/\delta))$$

Theorm: Let H be the hypothesis class over a domain X with a finite VC dimencion d. For every $\epsilon, \delta \in (0, 1)$, for any sampling distribution D and any time labelling f : X $\rightarrow \{+1, -1\}$,

$$P_{S \sim D^n}[\exists h \in H : E_S(h, f) = 0 \wedge E_D(h, f) > \epsilon] \leq \delta$$

whenever

$$n \geq 4/\epsilon(2d\,log(1/\epsilon) + log(2/\delta))$$

Theorm (Equivalently): A hypothesis class H with a finite VC-dimension d is PAC learnable with sample complexity

$$S_H(\epsilon, \delta) \leq 4/\epsilon(2d\,log(1/\epsilon) + log(2/\delta))$$

$$\in O(1/\epsilon(d\,log(1/\epsilon) + log(1/\delta)))$$

3 Sample Complexity Lower Bounds

→ CONVERSE: VC-dimension(H) = $\infty \Rightarrow$ Not PAC learnable?
(Answer is YES)
Given H, $\forall D, \forall f, \forall \epsilon, \forall \delta$
Claim: $S_H(1/2, 1/2) \geq 1/2\,VC - Dim(H)$
given H : any hypothesis class of VC dimension d.
given X : Domain of H
f(x) = +1 $\forall x \in X$
$T = \{x_1, ..., x_d\} \subseteq X$: A set shattered by H.
Clever Choice:
D : P(x) = {1/d $\forall x \in T$, and 0 $\forall x \in X\ T$ }
$\epsilon, \delta = 1/2$
Let n < $d/2$ and $S \sim D^n$
=> then, at least half the points in T are not in S.
=> $\exists h : E_S(h, f) = 0 \wedge E_D(h, f) > 1/2$
=> $P_{S \sim D^n}[\exists h : E_S(h, f) = 0 \wedge E_D(h, f) > \epsilon] = 1 > \delta(i.e.\,1/2)$
Hence $S_H(1/2, 1/2) \geq d/2$. (Note: S_H is decreasing in δ and ϵ so it is no better for smaller δ and ϵ)

→ Can we have a smarter algorithm?
Let 'A' be the smart algo.
$h^* = A(S)$ = Best choice among $\{h \in H : E_S(h, f) = 0\}$
We will beat it with same D but different f's.
Let f(x) = {+1 if $x \in X/T$, {+1 w.p. 1/2, -1 w.p. 1/2} if $x \in T$} ("Probabilistic Method") ($f \sim F$)
$n < d/2$
For any set S of at most n elements from T (At least 1/2 of T is outside S)

$$Expectation(E)_{f \sim F}[E_D(A(S), f)] = (1/2)|T/S| > 1/4$$

Hence $\exists f$ s.t. $\forall S$, $E_D(A(S), f) > 1/4 = \epsilon$
=> $P_{S \sim D^n}[E_D(A(S), f) > 1/4] = 1$
(Can a randomized algo. work? NO)

→ TIGHTNESS: For finite VC-dimension, can we reduce the sample complexity?
Final result -

$$S_H(h, f) = \Omega((d/\epsilon)log(1/\epsilon) + (1/\epsilon)log(1/\delta))$$

so tight upto the constraints
($log(1/\epsilon)$ term can be removed if we make 'smart algorithms')

Idea: $\delta_H(\epsilon, \delta) \geq \Omega(d/\epsilon)$
T $= \{x_1, ..., x_d\}$ shattered by H
$x_o \in X/T$
$D : P(x) = \{0\ if\ x \in X\backslash(T \cup \{x_0\}),\ 1 - 2\epsilon\ if\ x = x_0,\ 2\epsilon/d\ if\ x \in T\}$
Let $n < d/(8\epsilon)$

$$E_{S\sim D^n}[|S \cap T|] = nP_{x\sim D}[x \in T]$$

$$= n.2\epsilon < (d/(8\epsilon))2\epsilon = d/4$$

$$P_{S\sim D^n}[|S \cap T| \geq d/2] \leq 1/2\ \ (Markov\ Inequality)$$

$$P_{S\sim D^n}[|S \cap T| < d/2] > 1/2$$

$$|S \cap T| < d/2 \Rightarrow \exists h : E_S(h, f) = 0 \wedge E_D(h, f) > \epsilon$$

$$Hence\ P_{S\sim D^n}[\exists h : E_S(h, f) = 0 \wedge E_D(h, f) > \epsilon] > 1/2$$

$$i.e.\ S_H(\epsilon, 1/2) \geq d/(8\epsilon) = \Omega(d/\epsilon)$$

$$d/(8\epsilon) \leq S_H(\epsilon, 1/2) \leq (8d/\epsilon)log(1/\epsilon) + (4/\epsilon)log(2/(1/2)) = (8d/\epsilon)log(1/\epsilon) + 8/\epsilon$$

Rule of thumb : d/ϵ