Notes 12: Sample Size Bounds via VC dimension

Is $\mathcal{C}$ PAC-learnable?
How many samples are needed to learn $\mathcal{C}$? (perhaps with an inefficient algorithm)
If $\mathcal{C}$ is finite, and if confidence parameter δ is constant (e.g. $\delta = 1/100$)
then roughly $(\ln |\mathcal{C}|)/\varepsilon$ samples suffice (Consistent Hypothesis Algorithm)
What about lower bound?
What if $\mathcal{C}$ is infinite?

VC dimension gives almost tight answer!
Let $d = \mathrm{VCDim}(\mathcal{C})$
Any PAC learning algorithm for $\mathcal{C}$ must use $\Omega(d/\varepsilon)$ samples
if $\mathrm{VCDim}(\mathcal{C}) = \infty$, needs infinitely many samples (not PAC learnable)
Consistent Hypothesis Algorithm PAC-learns $\mathcal{C}$ with $m = O\left(\frac{1}{\varepsilon}\left(d \ln \frac{1}{\varepsilon} + \ln \frac{1}{\delta}\right)\right)$ samples
inefficient algorithm

$$C_1 \frac{d}{\varepsilon} \leqslant \#\text{samples to PAC learn (slowly)} \leqslant C_2 \frac{d \ln(1/\varepsilon) + \ln(1/\delta)}{\varepsilon}$$

---

1. Lower Bounds

Claim 1 (No Free Lunch)**.** *Let $d = \mathrm{VCDim}(\mathcal{C})$. Any PAC algorithm to learn $\mathcal{C}$ with $\delta = 1/10$ (say) must use $\geqslant d/2 = \Omega(d)$ samples on some distribution $\mathcal{D}$*

Proof. Some subset $S = \{x^1, \dots, x^d\}$ is shattered by $\mathcal{C}$
Every dichotomy $T \subseteq S$ is induced by some $c \in \mathcal{C}$
Idea: Every labeling is possible; $d/2$ seen samples give no information about unseen samples
$\mathcal{D}$ = uniform distribution on S
Pick one of the dichotomies T and some c inducing it (2^d of them) uniformly at random
If algorithm A gets $d/2$ samples and output hypothesis h

$$\mathop{\mathbb{E}}_{c}[\mathrm{err}_{\mathcal{D}}(h,c)] \geqslant \mathop{\mathbb{P}}_{x \sim \mathcal{D}}[x \text{ isn't among the } d/2 \text{ seen samples}] \mathop{\mathbb{P}}_{c}[h(x) \neq c(x)] = \frac{d/2}{d}\frac{1}{2} = \frac{1}{4}$$

$X \overset{\text{def}}{=} 1 - \mathrm{err}_{\mathcal{D}}(h,c)$ nonnegative random variable with $\mathbb{E}[X] \leqslant 3/4$
By averaging argument/Markov inequality,

$$\mathbb{P}[X \geqslant 7/8] \leqslant \mathbb{E}[X]/(7/8) \leqslant (3/4)/(7/8) = 6/7$$

i.e. $\mathbb{P}[\mathrm{err}_{\mathcal{D}}(h,c) \geqslant 1/8] \geqslant 1/7$ □

Markov inequality: For any nonnegative random variable X, any $t > 0$,

$$\mathbb{P}[X \geqslant t] \leqslant \mathbb{E}[X]/t$$

Reason: $$\mathbb{E}[X] = \mathbb{P}[X \geqslant t]\underbrace{\mathbb{E}[X \mid X \geqslant t]}_{\geqslant t} + \underbrace{\mathbb{P}[X < t]}_{\geqslant 0}\underbrace{\mathbb{E}[X \mid X < t]}_{\geqslant 0} \geqslant t\,\mathbb{P}[X \geqslant t]$$

The lower bound can be boosted to $\Omega(d/\varepsilon)$

Claim 2. *Let $d = \mathrm{VCDim}(\mathcal{C})$. Any PAC algorithm to learn $\mathcal{C}$ with $\delta = 1/10$ (say) must use $\Omega(d/\varepsilon)$ samples on some distribution $\mathcal{D}$*

Proof. Some subset $S = \{x^1, \dots, x^d\}$ is shattered by $\mathcal{C}$
$\mathcal{D}$ has weight $1 - 8\varepsilon$ on x^1 and weight $8\varepsilon/(d-1)$ on any of $x^2, \dots, x^d$
Idea: $x^2, \dots, x^d$ are rare: every $1/(8\varepsilon)$ sample is one of them; slows down learning by $\Omega(1/\varepsilon)$
Again, pick one of the dichotomies T and some c inducing it (2^d of them) uniformly at random
If algorithm A gets $\leqslant (d-1)/2$ of the rare samples (i.e. one of $x^2, \dots, x^d$)
then with prob. $\geqslant 1/7$, A has error $\geqslant 1/8$ under the uniform distribution over rare samples
rare samples have total weight 8ε, so A has error $\geqslant \varepsilon$ under $\mathcal{D}$

How likely will A get $\leqslant (d-1)/2$ of rare samples?

If A uses $\frac{d-1}{32\varepsilon} = \Omega(d/\varepsilon)$ samples

$\mathbb{E}[\#\text{rare samples}] = 8\varepsilon\frac{d-1}{32\varepsilon} = \frac{d-1}{4}$

$\mathbb{P}\left[\#\text{rare samples} \geqslant \frac{d-1}{2}\right] \leqslant e^{-(d-1)/12}$ (Chernoff; $pm = \frac{d-1}{4}, \gamma = 1$)

$\leqslant 1/100$ (say) when $d \geqslant 100$

Overall with prob. $\geqslant \frac{99}{100}\frac{1}{7} \geqslant \frac{1}{10}$, A outputs hypothesis h with error $\geqslant \varepsilon$ □

---

2. Upper bound

If $\text{VCDim}(\mathcal{C}) = d$, will show that $O\left(\frac{1}{\varepsilon}\left(d \ln \frac{1}{\varepsilon} + \ln \frac{1}{\delta}\right)\right)$ samples suffice to PAC-learn $\mathcal{C}$

Similar bound as Consistent Hypothesis analysis in notes09

$\ln |\mathcal{H}|$ replaced with $\text{VCDim}(\mathcal{C}) \ln \frac{1}{\varepsilon}$

Lower bound proof suggests too many dichotomies induced by $\mathcal{C}$ make future prediction difficult

Upper bound proof will show that when m is much bigger than d, not many dichotomies are possible

Will prove in two steps:

(1) When $m > d$, #dichotomies induced on m samples grow only polynomially, i.e. $O(m^d)$

(2) With few dichotomies, a small number of samples is likely representative and Consistent Hypothesis Algorithm works

---

Now measure #dichotomies on m samples as follows

Given subset of samples $S \subseteq X$

$$\Pi_{\mathcal{C}}(S) \stackrel{\text{def}}{=} \{\text{dichotomies induced on } S \text{ by } \mathcal{C}\} = \{c \cap S \mid c \in \mathcal{C}\}$$

e.g. $\mathcal{C} = \{\text{closed intervals}\}, S = \{1,2,3\} \subseteq X = \mathbb{R}$,

$$\Pi_{\mathcal{C}}(S) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}, \{1,2,3\}\} \qquad (\text{missing } \{1,3\})$$

Key definition: **Growth function / Shatter coefficient**

$$\Pi_{\mathcal{C}}(m) \stackrel{\text{def}}{=} \max \#\text{dichotomies induced on subset of } m \text{ samples} = \max\{\Pi_{\mathcal{C}}(S) \mid S \subseteq X, |S| = m\}$$

e.g. $\mathcal{C} = \{\text{closed intervals}\}$

$$\Pi_{\mathcal{C}}(1) = 2 \qquad \Pi_{\mathcal{C}}(2) = 4 \qquad \Pi_{\mathcal{C}}(3) = 7$$

Note: $\text{VCDim}(\mathcal{C}) \geqslant m \iff \Pi_{\mathcal{C}}(m) = 2^m$

$\Pi_{\mathcal{C}}(m)$ grows exponentially when $m \leqslant d$ (and that's why insufficient info to learn)

Next lecture: $\Pi_{\mathcal{C}}(m) \leqslant \left(\frac{em}{d}\right)^d$ grows polynomially in m when $m > d$ and d fixed