VC dimension

Sunday, October 17, 2021 8:56 PM

We have seen PAC and Mistake bound algorithms for many concept classes.

In the case of halfspaces there was a $\frac{1}{\gamma^2}$ dependence on the margin γ.

In fact, one can make this $\log \frac{1}{\gamma}$.

Suppose we predict majority of all surviving w.

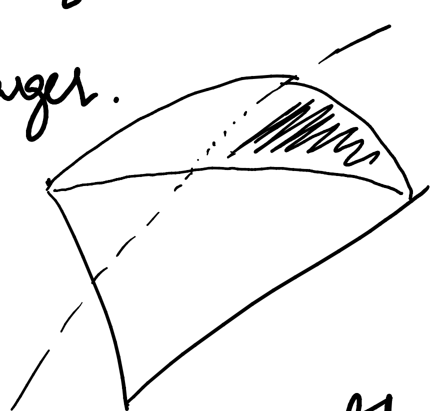
i.e. suppose after examples $x^1, x^2, \ldots x^l$,

we have $W = \{ w : w^T x^i \gtrless 0, \|w\| \le 1 \}$

as candidates and we consider which of

$$W \cap \{ w : w^T x^{l+1} \geq 0 \}, \quad W \cap \{ w : w^T x^{l+1} < 0 \}$$

is larger. Predict according to that.

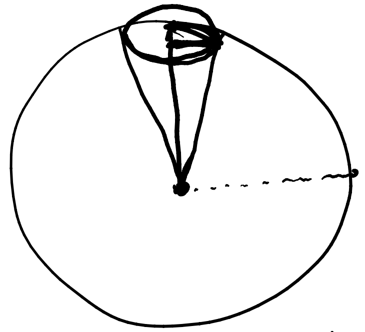


Then in each step we eliminate $\frac{1}{2}$ the volume.

$vol(W)$ starts at $vol(B)$.

vol(W) starts at vol(B).

At the end it is at least Vol(γ-cone)



$$\mathrm{Vol}(B) = \int_0^1 \left(\sqrt{1-t^2}\right)^{n-1} \mathrm{vol}(B^{n-1})\, dt$$

$$\mathrm{Vol}(\gamma\text{-cap}) = \int_{\sqrt{1-\gamma^2}}^1 \left(\sqrt{1-t^2}\right)^{n-1} \mathrm{vol}(B^{n-1})\, dt$$

$$\frac{\mathrm{Vol}(\gamma\text{-cap})}{\mathrm{Vol}(B)} = \frac{\int_{\sqrt{1-\gamma^2}}^1 (1-t^2)^{\frac{n-1}{2}} dt}{\int_0^1 (1-t^2)^{\frac{n-1}{2}} dt} \geq c \cdot \gamma^n.$$

$\therefore$ # mistakes $= O(n \log \frac{1}{\gamma})$.

How to estimate volume fraction?

We can sample the current W and take the majority vote of the sample. Even 1 sample suffices!

Alternatively we can use Linear Programming to find a feasible w for all constraints so far.

To bound the number of examples needed we can use a more general theory.

a more general theory.

VC-dimension. m points.

Concept class H.

How many distinct subsets of m points are defined by $h \in H$? $\quad H[m] \le m^{\text{VC-dim.}}$

More precisely: VC-dim (H) = max m s.t. $\exists$ m points that can be shattered, i.e. split in all possible ways by H.

e.g. intervals on a line $\quad$ VC-dim = 2

rectangles in 2-d $\quad$ VC-dim = 4

Halfspaces in $\mathbb{R}^d$ $\quad$ VC-dim = $d+1$

Thm 1. For a concept class H of VC-dim d, # distinct ways to split m points using $h \in H$ is $\le m^d$.

Thm 2. # examples needed to (ε,δ)-PAC learn H is $\le \frac{2}{\varepsilon}\left(\log(2H[2m]) + \log\frac{1}{\delta}\right)$.

is $\leq \frac{2}{\varepsilon}\left(\log(2\mathcal{H}[2m]) + \log\frac{1}{\delta}\right)$.

$= O\left(\frac{1}{\varepsilon}\left(d\log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right)\right)$.

Pf (Thm 1). We will show $\mathcal{H}[m] \leq \sum_{i=0}^{d}\binom{m}{i} = \binom{m}{\leq d}$.

Let S be a set of m points.

Induction on m. True for $m \leq d$.

Let $x \in S$. Consider $S \setminus \{x\}$.

By induction $\mathcal{H}(S\setminus\{x\}) \leq \binom{m-1}{\leq d}$.

Also note that

$$\binom{m}{\leq d} = \binom{m-1}{\leq d} + \binom{m-1}{\leq d-1}$$

So it suffices to show that

$$\mathcal{H}(S) - \mathcal{H}(S\setminus\{x\}) \leq \binom{m-1}{\leq d-1}$$

How can $\mathcal{H}(S)$ be larger? there must be labelings h and h' s.t. they agree on all
...

labelings h and h' s.t. they agree
points except x.

Let $T = \{ h \in \mathcal{H}(S) : h(x) = 1,\ h' \in \mathcal{H}(S) \}$.

Then we are interested in bonding $|T|$.

Let VC-dim $(T) = d'$. So $2^{d'}$ points can be shattered by T. But then $d'+1$ points can be shattered by H. So $d'+1 \le d$.

i.e. $d' \le d-1$.

Hence $\mathcal{H}(T) \le \binom{m-1}{\le d-1}$.

Pf (Thm 2). We find a hypothesis h_S that correctly classifies m points. We want to show that with prob $\geqslant 1-\delta$

$$\Pr_D \left(h_S(x) \neq h(x) \right) \le \varepsilon .$$

Let A be the complement of this event

Consider a different setting where we pick 2 subsets of size m, say

2 subsets of size ,

S and S'. Let B be the event that a hypothesis h has zero error on S and error $> \frac{\varepsilon}{2}$ on S'.

<u>Claim.</u> $Pr(B) \geq \frac{1}{2} Pr(A)$.

$$Pr(B) = Pr(A)\, Pr(B/A)$$

$Pr(B/A)$: Pr (h has error $\geq \frac{\varepsilon}{2}$ on m points given that it has error $\geq \varepsilon$ on D)

This is a simple chernoff bound.

$$Pr\left(\sum X_i - \mathbb{E}(\sum X_i) < \delta\, \mathbb{E}(\sum X_i) \right) \leq e^{-\frac{\delta^2 \mathbb{E}(\sum X_i)}{2}}.$$

$$Pr\left(\sum X_i < \frac{\varepsilon}{2} m \right) \leq e^{-\frac{\varepsilon m}{8}}.$$

$$m \geq \frac{8}{\varepsilon} \implies Pr(B/A) \geq \frac{1}{2}$$

So we want to show $Pr(B) \leq \frac{\delta}{2}$.

For this we pick 2m points, S'' partition them randomly into two subsets S, S' of m points.

randomly into two subsets S, S' of m points.

Then we want to bound $Pr(err_h(S)=0, err_h(S') > \frac{\varepsilon}{2})$.

Pair up the $2m$ points $(a_1, b_1), \ldots (a_m, b_m)$.

Fix hypothesis h.

if h makes error on both a_i and b_i,

then $Pr = 0$. (since no errors allowed on S).

Also at least $\frac{\varepsilon m}{2}$ indices i must make an error.

So $Pr(\text{all } \frac{\varepsilon m}{2} \text{ errors fall in } S') \leq \frac{1}{2^{\varepsilon m/2}}$.

$\#$ possible $h \leq \mathcal{H}(2m)$

$\therefore$ suffices to have $2^{-\varepsilon m/2} \mathcal{H}[2m] \leq \frac{\delta}{2}$

i.e. $m \geq \frac{2}{\varepsilon}\left(\log 2H[2m] + \log\frac{1}{\delta}\right)$.

<u>Chernoff</u>

$X = \Sigma X_i$ independent 0/1

$$Pr(X \geq (1+\delta)\mathbb{E}(X)) < e^{-\frac{\delta^2}{2+\delta}\mathbb{E}(X)}$$

$$Pr(X \leq (1-\delta)\mathbb{E}(X)) < e^{-\frac{\delta^2 \mathbb{E}(X)}{2}}$$

$Pr(X \leq (1-\delta)\mathbb{E}(X)) < e$

Hoeffding $a \leq X_i \leq b$

$$Pr(X \geqslant \mathbb{E}(X) + t) < e^{-\frac{2t^2}{n(b-a)^2}}$$

$$Pr(X \leqslant \mathbb{E}(X) - t) < e^{-\frac{2t^2}{n(b-a)^2}}$$

Pf (Thm 3). pairs (a_i, b_i) randomly allocate to S, S'.

$$|err_h(S) - err_h(S')| \geqslant \frac{\varepsilon}{2}$$

$Pr(S'$ gets $\frac{\varepsilon m}{2}$ more than $S)$

$$X_i = \begin{cases} 1 & \text{if } S' \\ -1 & \text{if } S \end{cases} \qquad \mathbb{E}(\Sigma X_i) = 0$$

$$Pr\left(\Sigma X_i \geqslant \frac{\varepsilon m}{2}\right) < e^{-2 \cdot \frac{\varepsilon^2 m}{4 \cdot 4}}$$

$$= e^{-\frac{\varepsilon^2 m}{8}}.$$

$$e^{-\frac{\varepsilon^2 m}{8}} \cdot \mathcal{H}(2m) \leq \frac{\delta}{2}$$

$$\Rightarrow m \geq \frac{8}{\varepsilon^2}\left(\log 2\mathcal{H}(2m) + \log \frac{1}{\delta}\right)$$

suffices.

VC-dim d : $m = O\left(\frac{1}{\varepsilon^2}\left(d \log \frac{1}{\varepsilon} + \log \frac{1}{\delta}\right)\right)$

suffices.