PACS Part 2, Lecture 7

PAC Learning

- PAC = probably approximately correct
- let's say we want to learn an interval $[a, b]$ based on samples from a distribution $\mathcal{D}$ on $\mathbb{R}$
- we won't be able to learn it exactly using a finite number of samples
- one sample = a point $x \sim \mathcal{D}$ and the information whether $x \in [a, b]$
- want measure of error of approximation of learned interval $[c, d]$:

$$\mathbb{P}(x \in [a, b]\Delta[c, d]) < \varepsilon$$

- optimum is $\varepsilon = 1$ in general, based on which samples we see
- but we can wish for an approximation error of ε of the learned interval with probability $\geq 1 - \delta$ (Monte Carlo algorithm)
- it turns out we only need $O(\frac{1}{\varepsilon} \log \frac{1}{\delta})$ samples to do just that

Sample Complexity

- a candidate interval $[c, d]$ can be discarded if we sample an $x \in [a, b]\Delta[c, d]$
- the number of samples to eliminate a single $[c, d]$ with $\mathbb{P}(x \in [a, b]\Delta[c, d]) \geq \varepsilon$ with probability $\geq 1 - \delta$ can be bounded by $m = \frac{1}{\varepsilon} \log \frac{1}{\delta}$
- application of the union bound gives $m = \frac{1}{\varepsilon} \log \frac{k}{\delta}$ samples to eliminate k such intervals
- however, we want to eliminate *all* of them simultaneously

VC Dimension

- VC = Vapnik–Chervonenkis
- A *range space* is a pair $(X, \mathcal{R})$ where X is a set and $\mathcal{R}$ is a set of subsets of X.
- If $S \subseteq X$, then $(S, \mathcal{R}_S)$ with $\mathcal{R}_S = \{R \cap S \mid R \in \mathcal{R}\}$ is also a range space. It is called the *projection* on S.
- A set $S \subseteq X$ is *shattered* by $\mathcal{R}$ if $|\mathcal{R}_S| = 2^{|S|}$.
- The *VC dimension* of $(X, \mathcal{R})$ is the maximum cardinality of a set $S \subseteq X$ that is shattered by $\mathcal{R}$.
- Example: $X = \mathbb{R}$, $\mathcal{R} = \{[a, b] \mid a, b \in \mathbb{R}\}$ has VC dimension 2

- Example: $X = \mathbb{R}^2$, $\mathcal{R} = \{C \subseteq X \mid C \text{ is convex}\}$ has infinite VC dimension (take n points on a circle)
- Example: $X = \{0,1\}$, $\mathcal{R} = \{f\colon X \to \{0,1\} \mid f \text{ is monotone}\}$ has VC dimension n (take the n points with exactly one zero; shatter by doing the AND of the never-zero variables)

Growth Function

- $\mathcal{G}(d,n) = \sum_{i=0}^{d} \binom{n}{i}$
- If $(X, \mathcal{R})$ is a range space of VC dimension d with $|X| = n$, then $|\mathcal{R}| \leq \mathcal{G}(n,d)$.

  Proof: By induction on d and n. The base cases with $d = 0$ or $n = 0$ hold since $\mathcal{G}(d,n) = 1$ in this case. For the induction step, assume that the claim holds for the pairs $(d-1, n-1)$ and $(d, n-1)$. Choose some $x \in X$ and consider the range spaces

  $$\begin{aligned}\mathcal{R}_1 &= \{R \setminus \{x\} \mid R \in \mathcal{R}\}\\ \mathcal{R}_2 &= \{R \setminus \{x\} \mid R \cup \{x\} \in \mathcal{R} \wedge R \setminus \{x\} \in \mathcal{R}\}\end{aligned}$$

  on the point set $X \setminus \{x\}$. We have $|\mathcal{R}| = |\mathcal{R}_1| + |\mathcal{R}_2|$. Further, the VC dimension of $(X \setminus \{x\}, \mathcal{R}_1)$ is $\leq d$ and that of $(X \setminus \{x\}, \mathcal{R}_2)$ is $\leq d-1$. We thus calculate:

  $$\begin{aligned}|\mathcal{R}| = |\mathcal{R}_1| + |\mathcal{R}_2| &\leq \mathcal{G}(d, n-1) + \mathcal{G}(d-1, n-1)\\ &= \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} = \sum_{i=0}^{d} \binom{n}{i} = \mathcal{G}(d,n)\end{aligned}$$

Component Bounds

- we can deduce bounds on the VC dimension of set-theoretic constructions from that for their components
- for example $\mathcal{R}^{\cup} = \{R_1 \cup R_2 \mid R_1 \in \mathcal{R}^1\,,\ R_2 \in \mathcal{R}^2\}$ and $\mathcal{R}^{\cap} = \{R_1 \cap R_2 \mid R_1 \in \mathcal{R}^1\,,\ R_2 \in \mathcal{R}^2\}$ both have VC dimension $O(d)$ if the components have VC dimension $\leq d$
- more generally: $\mathcal{R}^f = \{f(R_1, \ldots, R_k) \mid R_1 \in \mathcal{R}^1, \ldots, R_k \in \mathcal{R}^k\}$ has VC dimension $O(kd \log k)$
- a proof for the somewhat weaker bound $O(kd \log kd)$:

  Let t be the VC dimension of $(X, \mathcal{R}^f)$, and let $Y \subseteq X$ be shattered by $\mathcal{R}^f$ with $|Y| = t$. We have $|\mathcal{R}^i_Y| \leq \mathcal{G}(d,t) \leq t^d$, which implies:

  $$2^t = |\mathcal{R}^f_Y| \leq \prod_{i=1}^{k} |\mathcal{R}^i_Y| \leq t^{kd}$$

We will show $t < x \log x$ where $x = 2(kd+1)/\log 2$. Assume by contradiction that $t \geq x \log x$. Then

$$\frac{2t}{\log t} \geq \frac{2x \log x}{\log t} \geq \frac{2x \log x}{2 \log x} = x = \frac{2(kd+1)}{\log 2}$$

and thus $t \geq (kd+1)\log_2 t$ and $2^t \geq t^{kd+1} > t^{kd}$, a contradiction.

ε-Nets

- a set $N \subseteq X$ is an ε-*net* if every $R \in \mathcal{R}$ with $\mathbb{P}(R) \geq \varepsilon$ contains one point of N
- if $(X, \mathcal{R})$ has VC dimension d, then there is an

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

  such that a random sample of size m is an ε-net with probability $\geq 1 - \delta$

  Proof: Let M and T be two i.i.d. sets of m samples each. Let E_1 be the event that M is not an ε-net and let E_2 be the following event:

$$E_2\colon \quad \exists R \in \mathcal{R}\colon \mathbb{P}(R) \geq \varepsilon \wedge R \cap M = \emptyset \wedge |R \cap T| \geq \frac{\varepsilon m}{2}$$

  We have $\mathbb{P}(E_1) \leq 2\mathbb{P}(E_2)$ if $m \geq 8/\varepsilon$: Let $R' \in \mathcal{R}$ according to E_1. Then

$$\begin{aligned}\frac{\mathbb{P}(E_2)}{\mathbb{P}(E_1)} &= \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_1)} = \mathbb{P}(E_2 \mid E_1) \geq \mathbb{P}(|T \cap R'| \geq \varepsilon m/2) \\ &\geq 1 - e^{-\varepsilon m/8} \geq \frac{1}{2}\end{aligned}$$

  by the Chernoff bound.

  Defining the event

$$E_2'\colon \quad \exists R \in \mathcal{R}\colon R \cap M = \emptyset \wedge |R \cap T| \geq \frac{\varepsilon m}{2}$$

  we have $\mathbb{P}(E_2) \leq \mathbb{P}(E_2') \leq (2m)^d 2^{-\varepsilon m/2}$: Setting $k = \lceil \varepsilon m/2 \rceil$, the probability for some $R \in \mathcal{R}$ to have $R \cap M = \emptyset$ and $|R \cap T| \geq k$ is at most:

$$\begin{aligned}\mathbb{P}(M \cap R = \emptyset \mid |R \cap (M \cup T)| \geq k) &= \frac{\binom{2m-k}{m}}{\binom{2m}{m}} = \frac{(2m-k)!m!}{(2m)!(m-k)!} \\ &\leq 2^{-\varepsilon m/2}\end{aligned}$$

  Since the $|\mathcal{R}_{M \cup T}| \leq \mathcal{G}(2m, d) \leq (2m)^d$, we get the claimed inequality by the union bound.