25. Generalization Error in Classification

Wednesday, November 3, 2021 10:52 PM

Last time: • VC-dimension and shattering functions

Today: • Combining VC-dim of multiple set systems

• Applying VC-dim to understand generalization error.

We can think of a set system $\mathcal{H}$ as corresponding to some concept class, e.g. color. Combining set systems lets us combine together multiple concepts.

Lemma 5.11 Suppose $(X, \mathcal{H}_1)$ and $(X, \mathcal{H}_2)$ are set systems on the same X.

Then $\pi_{\mathcal{H}_1 \cap \mathcal{H}_2}(n) \le \pi_{\mathcal{H}_1}(n) \cdot \pi_{\mathcal{H}_2}(n)$, where $\mathcal{H}_1 \cap \mathcal{H}_2 = \{h_1 \cap h_2 \mid h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$.

proof. Let $A \subseteq X$, $|A| = n$. Let $S = \{A \cap h \mid h \in \mathcal{H}_1 \cap \mathcal{H}_2\}$.

By definition, $S = \{A \cap (h_1 \cap h_2) \mid h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$

$$\Rightarrow S = \{(A \cap h_1) \cap (A \cap h_2) \mid h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$$

$$\Rightarrow |S| \le |\{A \cap h_1 \mid h_1 \in \mathcal{H}_1\}| \, |\{A \cap h_2 \mid h_2 \in \mathcal{H}_2\}|$$

Choose A s.t. $|S| = \pi_{\mathcal{H}_1 \cap \mathcal{H}_2}(n)$.

Then $\pi_{\mathcal{H}_1 \cap \mathcal{H}_2}(n) \le \pi_{\mathcal{H}_1}(n) \pi_{\mathcal{H}_2}(n)$. $\square$

This allows us to take the Boolean AND of concepts.

i.e. if $X = \mathbb{R}^d$, and $\mathcal{H}_1 = \{\text{half-spaces}\}$ and $\mathcal{H}_2 = \{\text{half-spaces}\}$

$$\mathcal{H}_1 \cap \mathcal{H}_2 = \{\text{intersection of two half-spaces}\}$$

$$= \{\text{half-space 1 AND half-space 2}\}$$

Can extend to Boolean ANDs of many concepts.

Define Given k concepts $h_1, \dots, h_k \in \mathcal{H}$ and a Boolean function f, define the set

$$\text{comb}_f(h_1, \dots, h_k) = \{x \in X \mid f(h_1(x), \dots, h_k(x)) = 1\},$$

where $h_i(x) = 1$ iff $x \in h_i$.

Ex f is the AND function $\Rightarrow$ $\text{comb}_f(h_1, \dots, h_k) = \{x \in X \mid \prod_i h_i(x) = 1\}$

Ex. f is the majority-vote function $\Rightarrow$ $\text{comb}_f(h_1, \dots, h_k) = \{x \in X \mid \lfloor \frac{\sum h_i(x)}{k} + 0.5 \rfloor = 1\}$

Define: $COMB_{f,k}(\mathcal{H}) = \{comb_f(h_1, \dots, h_k) \mid h_i \in \mathcal{H}\}$ a new concept class.

Lemma 5.12 For any Boolean function f, hypothesis class $\mathcal{H}$, integer k,

$$\pi_{COMB_{f,k}(\mathcal{H})}(n) \le \pi_{\mathcal{H}}(n)^k.$$

proof. Same reasoning as 5.11.

Theorem 5.13 If $VC(\mathcal{H}) = V$, then $\forall$ Boolean function f and integer k,

$$VC(COMB_{f,k}(\mathcal{H})) = O(kV \log(kV))$$

proof. Let $n = VC(COMB_{f,k}(\mathcal{H}))$ We use n because it'll be the size of a shattered set.

By def, $\exists$ set S of n pts shattered by $COMB_{f,k}(\mathcal{H})$.

By Sauer's lemma, $\pi_{\mathcal{H}}(n) \le \binom{n}{\le V} \le n^V$, so there are at most n^V ways of partitioning S using sets in $\mathcal{H}$.

But each set in $COMB_{f,k}(\mathcal{H})$ is determined by k sets in $\mathcal{H}$, so there are at most $(n^V)^k = n^{Vk}$ ways of partitioning pts using $COMB_{f,k}(\mathcal{H})$

Since S is shattered, must have $2^n \le n^{kV} \Rightarrow n \le kV \log_2(n)$.

If $n \ge 16$, $\log_2(n) \le \sqrt{n} \Rightarrow kV \log_2(n) \le kV\sqrt{n}$

$$\Rightarrow n \le kV\sqrt{n} \Rightarrow n \le (kV)^2$$

$$\Rightarrow \Rightarrow n \le kV \log_2 (kV)^2 \le 2kV \log_2 kV.$$

(Key Thm)

Theorem 5.14 Let $(X, \mathcal{H})$ be a set system, D a prob dist over X, and let n be an integer satisfying $n \ge \frac{8}{\varepsilon}$ and

$$n \ge \frac{2}{\varepsilon}\left[\log_2 2\pi_{\mathcal{H}}(2n) + \log_2 \frac{1}{\delta}\right].$$

Let S_1 consist of n pts drawn from D, possibly with repetition.

With prob. $\ge 1-\delta$, every set in $\mathcal{H}$ of prob. mass $> \varepsilon$ intersects S_1.

Note: If $VC(\mathcal{H}) = d < \infty$, $\log(\pi_{\mathcal{H}}(2n)) = O(d \log n)$ by Sauer, and we can achieve an inequality $n \ge a \log n$ $(a \ge 4)$ by $n \ge ca \log a$ for some constant c.

proof. Let event $A = \{\exists h \in \mathcal{H}$ with $\mu(h) \ge \varepsilon$ s.t. $h \cap S_1 = \emptyset\}$

Draw a second set S_2 of n pts from D.

Let event $B = \{\exists h \in \mathcal{H}$ with $h \cap S_1 = \emptyset$ but $|h \cap S_2| \geq \frac{\varepsilon}{2} \cdot n\}$

If $h \cap S_1 = \emptyset$, and $\mu(h) \geq \varepsilon$, then $\mathbb{E}|h \cap S_2| = \mathbb{E} \sum_{x \in S_2} \mathbb{1}_{\{x \in h\}} \geq n\varepsilon$.

Also, $Var(|h \cap S_2|) = n \, Var(\mathbb{1}_{\{x \in h\}}) = n\,\varepsilon(1-\varepsilon) \leq n\varepsilon$.

By Chebyshev, $Prob\left(|h \cap S_2| \geq \frac{\varepsilon n}{2}\right) \leq n\varepsilon \cdot \left(\frac{2}{n\varepsilon}\right)^2 = \frac{4}{n\varepsilon} \leq \frac{1}{2}$.

$n \geq \frac{8}{\varepsilon}$

$\Rightarrow Prob(B|A) \geq \frac{1}{2} \qquad \Rightarrow Prob(B) \geq \frac{1}{2} Prob(A)$

and this only counts when its the same h for A & B.

Thus, to prove $Prob(A) \leq \delta$, it would suffice to prove $Prob(B) \leq \frac{\delta}{2}$.

Consider drawing S_3 of $2n$ pts and randomly partitioning into lists S_1 and S_2. Clearly yields same prob. dist.

Let's hold off on partitioning S_3.

Note that $|\{S_3 \cap h \mid h \in \mathcal{H}\}| \leq \pi_{\mathcal{H}}(2n)$ (even if $|\mathcal{H}| = \infty$)

So $Prob(B) \leq \sum_{h' \in \{S_3 \cap h \mid h \in \mathcal{H}\}} Prob\left(|S_1 \cap h'| = 0 \text{ AND } |S_2 \cap h'| \geq \frac{\varepsilon}{2} n\right)$

$\leq \pi_{\mathcal{H}}(2n) \cdot Prob\left(|S_1 \cap h'| = 0 \text{ AND } |S_2 \cap h'| \geq \frac{\varepsilon}{2} n\right) \quad \forall h'$.

So to prove $Prob(B) \leq \frac{\delta}{2}$, suffices to show

$$Prob\left(|S_1 \cap h'| = 0 \text{ AND } |S_2 \cap h'| \geq \frac{\varepsilon}{2} n\right) \leq \frac{\delta}{2\pi_{\mathcal{H}}(2n)}$$

Note $|h'| \geq \frac{\varepsilon}{2} n$, because otherwise $|S_2 \cap h'| < \frac{\varepsilon}{2} n$.

Thus, $\underset{\substack{\text{random}\\\text{partition}\\\text{of } S_3}}{Prob}(|S_1 \cap h'| = 0) \leq \left(\frac{1}{2}\right)^{\frac{\varepsilon n}{2}}$ because each item in h' has a $\frac{1}{2}$ chance of falling in S_1 vs. S_2

$$= 2^{-\frac{\varepsilon n}{2}} \leq 2^{-\log_2 2\pi_{\mathcal{H}}(2n) + \log_2 \delta} = \frac{\delta}{2\pi_{\mathcal{H}}(2n)}$$

$\boxtimes$

Proof technique where we picked S_1 & S_2 two different ways is known as "double sampling". We postpone random choices until later like in percolation theory proofs.

Formalizing the classification generalization problem

Formalizing the classification generalization problem

Given a prob. dist. D over space X, we receive training set S drawn from D.
We want to predict well on new points from D.

Let $c^* \subseteq X$ be a target concept (e.g. spam emails)

We want hypothesis $h \subseteq X$ s.t. the symmetric difference $h \Delta c^*$ is minimized.

Define true error of h $\quad err_D(h) = \mu(h \Delta c^*)$

training error of h $\quad err_S(h) = \frac{|S \cap (h \Delta c^*)|}{|S|}$

Unfortunately, minimizing $err_S(h)$ may not minimize $err_D(h)$ because of overfitting.

Instead, we turn to a restricted class of hypotheses $\mathcal{H} \subseteq 2^X$, i.e. set systems $(X, \mathcal{H})$.

When can we hope to generalize + not overfit? One condition is that sample size is large compared to VC-dim.

Let $\mathcal{H}' = \{h \Delta c^* \mid h \in \mathcal{H}\}$ be the collection of error regions of hypotheses in $\mathcal{H}$ }

Lemma: $\mathcal{H}$ and $\mathcal{H}'$ have the same VC-dim + shatter function.

proof. Exercise for reader. (also in MAT1801-2020 lecture 9.5)

Thm 5.15 (sample bound): For any class $\mathcal{H}$ and distribution D, if a training sample of size S is drawn from D of size

$$n \geq \frac{2}{\varepsilon}\left[\log\left(2\pi_{\mathcal{H}}(2n)\right) + \log\frac{1}{\delta}\right],$$

then w.p. $\geq 1-\delta$, every $h \in \mathcal{H}$ with training error $err_S(h) = 0$ has $err_D(h) < \varepsilon$.

proof. Apply Thm 5.14 to $\mathcal{H}' = \{h \Delta c^* \mid h \in \mathcal{H}\}$.

Thm 5.16 (growth function uniform convergence)

If training sample S has size

$$n \geq \frac{8}{\varepsilon^2}\left[\log\left(2\pi_{\mathcal{H}}(2n)\right) + \log\frac{1}{\delta}\right],$$

then w.p. $1-\delta$, every $h \in \mathcal{H}$ will have $|err_S(h) - err_D(h)| \leq \varepsilon$.

proof. Similar to 5.14 and 5.15, apply Chernoff-Hoeffding bounds.

Corollary 5.17 (from 5.16) For any class $\mathcal{H}$, dist D, a training sample S of size

$$O\left(\frac{1}{\varepsilon}\left[VC(\mathcal{H}) \log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right]\right)$$

is sufficient to ensure w.p. $1-\delta$ that every $h \in \mathcal{H}$ with $err_S(h) = 0$ has $err_D(h) < \varepsilon$.

VC-dim is one measure of the complexity of a set system, which allows proving generalization guarantees. There are others, such as Shannon entropy or Rademacher complexity (how well a concept class can fit random noise).

These types of guarantees give us hope that we can train a ML algorithm on a small sample of data and make useful predictions elsewhere