VC GENERALIZATION BOUND

Matthieu Bloch

February 19, 2019

LOGISTICS

Lecture slides and notes

- Report typos and errors on Piazza (thank you!)

If you are considering dropping the class, please send an email to Dr. Bloch

Problem set #1

- Solutions released, check it out
- Grading in progress

Problem set #2

- Be a bit more patient, it's coming asap

Midterm

- March 5, 2019 (Withdrawal deadline on March 13, 2019)
- 75 minutes, in class
- Two/three problems testing understanding and applications of concepts
- Open notes

RECAP: DICHOTOMIES AND GROWTH FUNCTION

Definition (Dichotomy)

For a dataset $\mathcal{D} \triangleq \{\mathbf{x}_i\}_{i=1}^N$ and set of hypotheses $\mathcal{H}$, the set of *dichotomies* generated by $\mathcal{H}$ on $\mathcal{D}$ is

$$\mathcal{H}(\{\mathbf{x}_i\}_{i=1}^N) \triangleq \{\{h(\mathbf{x}_i)\}_{i=1}^N : h \in \mathcal{H}\}$$

By definition $|\mathcal{H}(\{\mathbf{x}_i\}_{i=1}^N)| \leq 2^N$ and in general $|\mathcal{H}(\{\mathbf{x}_i\}_{i=1}^N)| \ll |\mathcal{H}|$

Definition (Growth function)

For a set of hypotheses $\mathcal{H}$, the *growth function* of $\mathcal{H}$ is

$$m_{\mathcal{H}}(N) \triangleq \max_{\{\mathbf{x}_i\}_{i=1}^N} |\mathcal{H}(\{\mathbf{x}_i\}_{i=1}^N)|$$

The growth function does *not* depend on the datapoints $\{\mathbf{x}_i\}_{i=1}^N$

The growth function in bounded $m_{\mathcal{H}}(N) \leq 2^N$

RECAP: BREAK POINT

Linear classifiers: $\mathcal{H} \triangleq \{h : \mathbb{R}^2 \to \{\pm 1\} : \mathbf{x} \mapsto \operatorname{sgn}(\mathbf{w}^\intercal \mathbf{x} + b) | \mathbf{w} \in \mathbb{R}^2, b \in \mathbb{R}\}$

- $m_{\mathcal{H}}(3) = 8$
- $m_{\mathcal{H}}(4) = 14 < 2^4$

Definition (Shattering)

If $\mathcal{H}$ can generate all dichotomies on $\{\mathbf{x}_i\}_{i=1}^N$, we say that $\mathcal{H}$ *shatters* $\{\mathbf{x}_i\}_{i=1}^N$

Definition (Break point)

If no data set of size k can be shattered by $\mathcal{H}$, then k is a break point for $\mathcal{H}$

The break point for linear classifiers is 4

Proposition.

If there exists *any* break point for $\mathcal{H}$, then $m_{\mathcal{H}}(N)$ is polynomial in N

If there is no break point for $\mathcal{H}$, then $m_{\mathcal{H}}(N) = 2^N$

VC GENERALIZATION BOUND

Consider our learning problem from Lecture 2

Proposition (VC bound)

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|R(h)-\widehat{R}_N(h)\right|>\epsilon\right)\le 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}$$

Compare this with our previous generalization bound that assumed $|\mathcal{H}|<\infty$

$$\mathbb{P}\left(\max_{h\in\mathcal{H}}\left|R(h)-\widehat{R}_N(h)\right|>\epsilon\right)\le 2|\mathcal{H}|e^{-2\epsilon^2 N}$$

- We replace the $\max$ by $\sup$ and $|\mathcal{H}|$ by $m_{\mathcal{H}}(2N)$
- We can now handle *infinite* hypothesis classes!

With probability at least $1-\delta$

$$R(h^*)\le \widehat{R}_N(h^*)+\sqrt{\frac{8}{N}\left(\log m_{\mathcal{H}}(2N)+\log\frac{4}{\delta}\right)}$$

Key insight behind proof is how to relate $\sup_{h\in\mathcal{H}}$ to $\max_{h\in\mathcal{H}'}$ with $\mathcal{H}'\subset\mathcal{H}$ and $|\mathcal{H}'|<\infty$

Approach developed by *Vapnik* and *Chervonenkis* in 1971

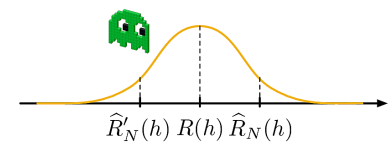
KEY INSIGHTS OF VC BOUND

The growth function $m_{\mathcal{H}}$ plays a role

- There may be infinitely many $h\in\mathcal{H}$, but they generate a finite number of unique dichotomies
- Hence, $\{\widehat{R}_N(h):h\in\mathcal{H}\}$ is *finite*
- Unfortunately $R(h)$ still potentialy takes infinitely many different values

Key insight: use a second *ghost* dataset of size N with empirical risk $\widehat{R}'_N(h)$

- Hope that we can squeeze $R(h)$ between $\widehat{R}'_N(h)$ and $\widehat{R}_N(h)$



We will try to relate $\mathbb{P}\left(\left|R(h)-\widehat{R}_N(h)\right|>\epsilon\right)$ to $\mathbb{P}\left(\left|\widehat{R}'_N(h)-\widehat{R}_N(h)\right|>\epsilon'\right)$ with $\epsilon'=f(\epsilon)$

- $\mathbb{P}\left(\left|\widehat{R}_N(h)-\widehat{R}'_N(h)\right|>\epsilon\right)$ only depends on the finite number of unique dichotomies

INTUITION

Assume that X,X' be i.i.d. random variables with *symmetric* distribution around their mean μ

- Let $\mathcal{A}\triangleq\{|X-\mu|>\epsilon\}$
- Let $\mathcal{B}\triangleq\{|X-X'|>\epsilon\}$

Lemma (Symmetric bound)

$$\mathbb{P}(\mathcal{A})\le 2\mathbb{P}(\mathcal{B})$$

If $X\triangleq\widehat{R}_N(h)$ and $X'\triangleq\widehat{R}'_N(h)$ had symmetric distributions, we would obtain

$$\mathbb{P}\left(\left|R(h)-\widehat{R}_N(h)\right|>\epsilon\right)\le 2\mathbb{P}\left(\left|\widehat{R}_N(h)-\widehat{R}'_N(h)\right|>\epsilon\right)$$

Not quite true, but close

PROOF OF VC BOUND

Lemma.

If $N\ge 4\epsilon^{-2}\ln 2$,

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|R(h)-\widehat{R}_N(h)\right|>\epsilon\right)\le 2\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\widehat{R}'_N(h)-\widehat{R}_N(h)\right|>\frac{\epsilon}{2}\right)$$

Lemma.

Let $\mathcal{S}\triangleq\{(\mathbf{x}_i,y_i)\}_{i=1}^{2N}$ be a dataset partitioned into two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ of N points. Assume that $\widehat{R}_N(h)$ is computed on $\mathcal{S}_1$ while $\widehat{R}'_N(h)$ is computed on $\mathcal{S}_2$.

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\widehat{R}'_N(h)-\widehat{R}_N(h)\right|>\frac{\epsilon}{2}\right)\le m_{\mathcal{H}}(2N)\sup_{\mathcal{S}_1,\mathcal{S}_2}\sup_{h\in\mathcal{H}}\mathbb{P}\left(\left|\widehat{R}'_N(h)-\widehat{R}_N(h)\right|\middle|\mathcal{S}_1,\mathcal{S}_2\right)$$

Lemma. For any $h\in\mathcal{H}$ and any partition $\mathcal{S}_1,\mathcal{S}_2$, we have

$$\mathbb{P}\left(\left|\widehat{R}'_N(h)-\widehat{R}_N(h)\right|\middle|\mathcal{S}_1,\mathcal{S}_2\right)\le 2e^{-\frac{1}{8}\epsilon^2 N}$$