Tuesday, 9/22/2020

EE 660

MACHINE LEARNING FROM SIGNALS: FOUNDATIONS AND METHODS

Prof. B. Keith Jenkins

Lecture 9

p. 1

Lecture 9	EE 660	Sep 22, 2020

Announcements

- Homework 3 is due Friday.
- Homework 1 has been graded
  - Your scores and any comments/markups are visible to you on D2L.

Today's Lecture

- Upper bounds on growth function
- VC generalization bound
  - Theorem
  - Interpretation
  - Implications
- Dataset methodology and generalization bounds (part 1)

Bounds on Growth Function $m_{\mathcal{H}}(N)$

We know that $m_{\mathcal{H}}(N) \leq 2^N$
and if k is a break point then $m_{\mathcal{H}}(k) < 2^k$.
We can state other upper bounds (sometimes more useful)

1. Theorem 2.4 If k is a break point for $\mathcal{H}$, then:

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i} \quad \forall N$$

Comment: $\binom{N}{i} = \dfrac{N!}{(N-i)!\,i!} = \dfrac{1}{i!}\left[N(N-1)\cdots(N-i+1)\right]$

polynomial in N? Yes. Degree? i.

$$= P_i(N)$$

$\sum_{i=0}^{k-1} \binom{N}{i}$? $i=k-1$ gives highest order polynomial

Σ is a sum of polynomials

$$\therefore \sum_{i=0}^{k-1} \binom{N}{i} = P_{k-1}(N), \text{ if } k \text{ is a break point.}$$

2. Let k_0 be the smallest break point of $\mathcal{H}$.

What is the relation $d_{vc}(\mathcal{H}) \leftrightarrow k_0$? $d_{vc}(\mathcal{H}) = k_0 - 1$

max # pts. $\mathcal{H}$ can shatter min. # pts $\mathcal{H}$ cannot shatter

One can show: $m_{\mathcal{H}}(N) \leq N^{d_{vc}} + 1$; d_{vc} is the order of the polynomial bound.

VC Generalization Bound

(1) Previously: $P\left[E_{out}(h_g) \le E_{\mathcal{D}}(h_g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}} \right] \ge 1-\delta$

(AML (2.1) again)

Now we have;

(2)

> **Theorem 2.5 – VC Generalization Bound**
>
> For any tolerance $\delta > 0$,
>
> $$E_{out}(h_g) \le E_{\mathcal{D}}(h_g) + \underbrace{\sqrt{\frac{8}{N} \ln \frac{4\, m_{\mathcal{H}}(2N)}{\delta}}}_{\varepsilon_{eff}}$$
>
> with probability $1-\delta$.

(AML (2.12))

Proof: in AML appendix (N.R.F.)

Comments

1. $m_{\mathcal{H}}(N) \le N^{d_{vc}} + 1 \implies m_{\mathcal{H}}(2N) \le (2N)^{d_{vc}} + 1$

(3) So: $\varepsilon_{eff} \le \sqrt{\frac{8}{N} \ln \frac{4\left[(2N)^{d_{vc}} + 1\right]}{\delta}} \triangleq \varepsilon_{vc}$

For $(2N)^{d_{vc}} \gg 1$:

$$\varepsilon_{vc} \approx \sqrt{\frac{8}{N}\left[\ln\left(4(2N)^{d_{vc}}\right) - \ln\delta\right]}$$

$$= \sqrt{\frac{8\ln 4}{N} + \frac{8\,d_{vc}}{N}\ln(2N) - \frac{8}{N}\ln\delta}$$

$$\Rightarrow \lim_{N\to\infty} \varepsilon_{vc} = \,? \; = 0.$$

$\varepsilon_{vc} \to 0$ asymptotically with increasing N.

If d_{vc} is finite, then a sufficiently large N exists that will provide good generalization from $E_{\mathcal{D}}$ to E_{out}.

> This proves that machine learning is feasible, even with an infinite hypothesis set ($M = |\mathcal{H}| = \infty$), if d_{vc} is finite

2. Using a test set to bound $E_{out}(h_g)$:

If $\mathcal{D}_{Test}$ has not been used to pick $\mathcal{H}$ or h_g, and if we use Eq. (1):

$$E_{out}(h_g) \le E_{\mathcal{D}_{Test}}(h_g) + \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}, \quad \text{with probability } 1-\delta,$$

then $M = ? = 1$!

3. Eq. (2) again: $E_{out}(h_g) \le E_{\mathcal{D}}(h_g) + \sqrt{\frac{8}{N}\ln\frac{4\, m_{\mathcal{H}}(2N)}{\delta}}$

using (3): $E_{out}(h_g) \le E_{\mathcal{D}}(h_g) + \sqrt{\frac{8}{N}\ln\frac{4\left[(2N)^{d_{vc}}+1\right]}{\delta}}$

d_{vc} is complexity of $\mathcal{H}$ (or model complexity)

N represents sample (dataset) complexity *

ε (ε_M, ε_{eff}, or ε_{vc}) is generalization error ($E_{\mathcal{D}} \rightarrow E_{out}$)

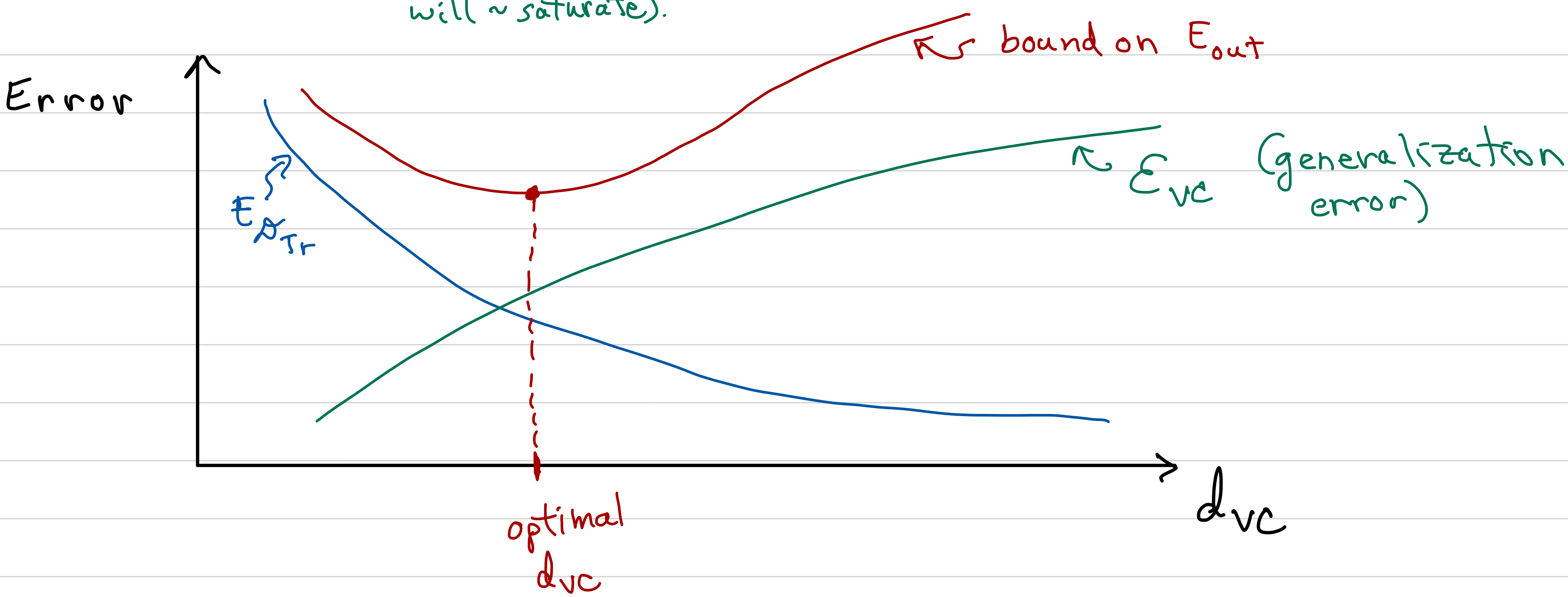
δ is our tolerance (degree of uncertainty we will accept)

* assuming drawn iid from $\mathcal{X}$ based on $p(\underline{x})$ (and $p(y|\underline{x})$ or $p(\underline{x}, y)$)

Dependences:

$$E_{out}(h_g) \le E_{\mathcal{D}}(h_g) + \sqrt{\frac{8}{N}\ln\frac{4[(2N)^{d_{vc}}+1]}{\delta}}$$

if $d_{vc}\downarrow$ then:	typ.	$\uparrow$	$\downarrow$
if $d_{vc}\uparrow$ then:		$\downarrow$	$\uparrow$
if $N\uparrow$ then:	maybe $\downarrow$ (at some large N, will ~ saturate).	maybe $\uparrow$, maybe ~const.	$\downarrow$



[Adapted from AML Fig. 2.3]

Review: 2 key bounds and their assumptions

(i) $E_{out}(h_g) \le E_{\mathcal{D}_0}(h_g) + \sqrt{\frac{8}{N} \ln \frac{4\left[(2N)^{d_{vc}}+1\right]}{\delta}}$ (or in terms of $m_{\mathcal{H}_0}(2N)$)

based on $\mathcal{H}_0$

Assumptions:

1. $\mathcal{D}_0$ and $\mathcal{H}_0$ must be consistent: i.e., $\mathcal{D}_0$ is used to choose h_g out of $\mathcal{H}_0$.
2. Info. in $\mathcal{D}_0$ cannot be used to construct $\mathcal{H}_0$.
3. $N = N_{\mathcal{D}_0}$

(ii) $E_{out}(h_g) \le E_{\mathcal{D}_{Test}}(h_g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$, $M = 1$.

Assumptions:

1. Info. in $\mathcal{D}_{Test}$ cannot influence choice of h_g in any way.
2. $N = N_{Test}$

And as always we assume data is drawn from $\mathcal{X}$ according to $p(\underline{x})$ i.i.d. (possibly also $p(y|\underline{x})$, or $p(\underline{x}, y)$, or $p(\underline{x}|y)$)