Announcements

- Homework 5 is due Tuesday.

Today's Lecture

- Interpreting the VC generalization bound
- Implications in dataset methodology

VC GENERALIZATION BOUND AND ITS INTERPRETATION

PLUG IN: $m_{\mathcal{H}}(2N) \leq (2N)^{d_{vc}} + 1$ TO $\varepsilon_{eff} \Rightarrow$

1. $$\varepsilon_{eff} \leq \sqrt{\frac{8}{N} \ln \frac{4\left[(2N)^{d_{vc}} + 1\right]}{\delta}} = \varepsilon_{VC}$$

FOR: $(2N)^{d_{vc}} >> 1$

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

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

$$\lim_{N\to\infty} \varepsilon_{VC} = 0.$$

$\Rightarrow$ THIS PROVES THAT LEARNING IS FEASIBLE, EVEN WITH INFINITE HYPOTHESIS SETS, 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 h_g OR $\mathcal{H}$, THEN:

$$E_{OUT}(h_g) \leq E_{Test}(h_g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$$

WITH PROBABILITY $1-\delta$.

→ USE $M = 1$.

3. Th'm 2.5 AGAIN: (using $\mathcal{E}_{VC}$)

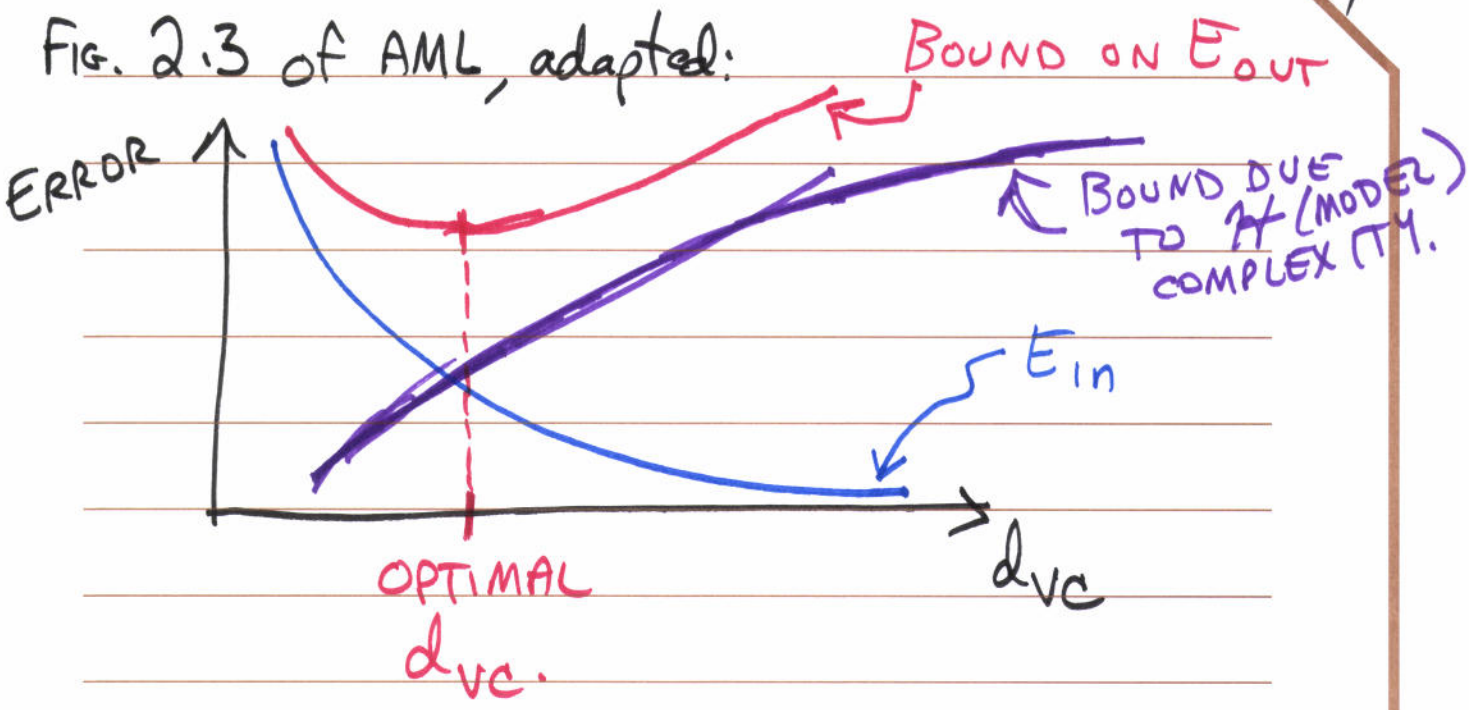
$$E_{OUT}(h_g) \leq E_{in}(h_g) + \sqrt{\frac{8}{N} \ln \frac{4[(2N)^{d_{VC}} + 1]}{\delta}}$$

d_{VC} ↓ typ. ↑

d_{VC} ↑ typ. ↓

BOUND DUE TO MODEL COMPLEXITY (FOR A GIVEN N, δ).

Fig. 2.3 of AML, adapted:



Implications in Dataset Methodology

Consider this scenario:

I. 1. Collect data and construct $\mathcal{D}$.

2. Do preliminary data analysis.
   - Visualize (or look at) data.
   - Plot some histograms, learn from them..

3. Preprocessing
   - ~~Normalize~~ Normalize
   - ⋮

4. Feature extraction.
   - Assess whether extracted features are useful for classification.

5. Do some preliminary trials on $\mathcal{D}$.
   - Varying dimension (#features)
   - Try a few hypothesis sets and learning algorithms.

6. SET UP AND RUN MODEL SELECTION AND LEARNING ALGORITHMS.

- DECIDE ON $\mathcal{H}$
- DIVIDE $\mathcal{D} \begin{cases} \mathcal{D}' \\ \mathcal{D}_{Test} \end{cases}$
- FOR CROSS-VALIDATION ON $\mathcal{D}'$ TO CHOOSE PARAMETERS, TRAIN, AND FIND h_g. (USING $\mathcal{D}' \begin{cases} \mathcal{D}_{Tr} \\ \mathcal{D}_{Val} \end{cases}$)

7. EVALUATE ITS PERFORMANCE:

(a) CALCULATE $E_{in}(h_g)$ USING $\mathcal{D}_{Tr}$ OR $\mathcal{D}_{Val}$
CALCULATE $\mathcal{E}_{VC}$ USING d_{VC} OF $\mathcal{H}$ (ALSO N, δ).
GET "ERROR BAR" ON $E_{out}(h_g)$.

(b) CALCULATE $E_{Test}(h_g)$ USING $\mathcal{D}_{Test}$
CALCULATE $\mathcal{E}_M$ USING $M = 1$
GET "ERROR BAR" ON $E_{OUT}(h_g)$.

ARE THESE $E_{OUT}(h_g)$ ERROR BARS VALID?

7(b) - No.

7(a) - No.

How to fix this?

For 7(b): Modify I to separate $\mathcal{D} \subset \mathcal{D}' \subset \mathcal{D}_{Test}$ after step 1.

Then use only $\mathcal{D}'$ in steps 2-6.

7(b) will be valid.