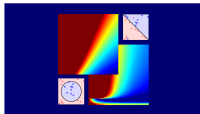


Machine Learning Foundations

(機器學習基石)



Lecture 7: The VC Dimension, Extended

Hsuan-Tien Lin (林軒田)

`htlin@csie.ntu.edu.tw`

Department of Computer Science
& Information Engineering

National Taiwan University
(國立台灣大學資訊工程系)



VC Dimension of Perceptrons

- 1D perceptrons (pos/neg rays): $d_{VC} = 2$
 - 2D perceptrons: $d_{VC} = 3$
 - d -D perceptrons: $d_{VC} = d + 1$
-
- 2D perceptrons **passing origin**: $d_{VC} = 2$ by $m_{\mathcal{H}}(N) = 2N$
 - d -D perceptrons **passing origin**: $d_{VC} = d$,
provable with linear independence/dependence arguments

proving d_{VC} **sometimes easier than**
deriving $m_{\mathcal{H}}(N)$ (but still hard)

Chain of Thought of VC Bound

- characterize $\mathcal{H}$ by **its growth function** $m_{\mathcal{H}}(N)$ in a worst-case scenario
- simplify/bound $m_{\mathcal{H}}(N)$ by a polynomial using the **maximum non-break point** $d_{\text{VC}}(\mathcal{H})$ as complexity measure of $\mathcal{H}$
- with **a high probability**,

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \underbrace{\sqrt{\frac{8}{N} \ln \left(\frac{4(2N)^{d_{\text{VC}}}}{\delta} \right)}}_{\Omega(N, \mathcal{H}, \delta)}$$

is VC bound still valid in deep learning age?

Validity of VC Bound for Deep Learning

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \underbrace{\sqrt{\frac{8}{N} \ln \left(\frac{4(2N)^{d_{\text{VC}}}}{\delta} \right)}}_{\Omega(N, \mathcal{H}, \delta)}$$

meaningful

- still **true**, as a mathematical theorem
- conceptually useful in connecting **model complexity** with **generalization performance**

not meaningful

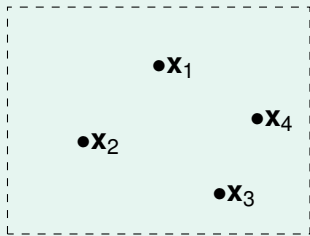
- vacuous, $\Omega \gg 1$ for reasonable N and $\mathcal{H}$, because of looseness
- new observations (e.g. double-descent) **not totally explained yet**

take the **philosophical message**, not the mathematical numbers

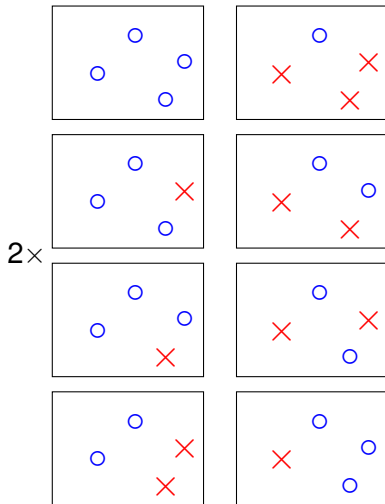
Another Complexity Measure: Average Minimum E_{in}

$$\mathcal{H} = \left\{ \text{all lines in } \mathbb{R}^2 \right\}$$

for four inputs $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$



$\min E_{in} = 0$ for 14 cases
 $\min E_{in} = 1/4$ for two cases
 $\text{avg} = 1/32$



Average Minimum $E_{\text{in}} \equiv$ Rademacher Complexity
for binary classification,

$$E_{\text{in}}(h) = \frac{1}{N} \sum_{n=1}^N \mathbb{I}[y_n \neq h(\mathbf{x}_n)] = \frac{1}{2N} \sum_{n=1}^N (1 - \underbrace{y_n h(\mathbf{x}_n)}_{\text{'correlation'}})$$

- data-dependent Rademacher complexity $\mathcal{R}_{\mathcal{H}}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$: **maximum 'correlation'** of $h \in \mathcal{H}$ to random $\mathbf{y} = \{-1, +1\}^N$
- **softer than growth function**, easily extended to **regression**
- with a high probability,

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \underbrace{2\mathcal{R}_{\mathcal{H}}(\mathbf{x}_1, \dots, \mathbf{x}_N)}_{\Omega} + 3\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2N}}$$

Rademacher complexity bound (here):
a **data-dependent** bound