MAC 0460 / 5832
Introduction to Machine Learning

11 – Is learning feasible ?

• Bias-variance tradeoff • Additional references •

IME/USP (19/05/2021)

Recap: **VC analysis**

VC inequality

$$\boxed{P\left(\left|E_{in}(g) - E_{out}(g)\right| > \varepsilon\right) \leq 4\, \underbrace{m_{\mathcal{H}}(2N)}_{\text{bound}}\, e^{-\frac{1}{8}\varepsilon^2 N}}$$

Learning is feasible when, for some finite N, with high probability

$$\underbrace{\left|\, E_{in} - E_{out} \,\right| < \varepsilon}$$

Recap: **VC analysis**

$$E_{out}(g) \leq E_{in}(g) + \Omega(N, \mathcal{H}, \delta)$$

In fact, what we would like to have is a small E_{out}

The above inequality shows that we need to be able to control E_{in} and Ω simultaneously

- E_{in} is related to fitting / approximation
- Ω is related to generalization

$$E_{out}(g) \leq E_{in}(g) + \Omega(N, \mathcal{H}, \delta)$$

Fitting / approximation

How well $\mathcal{H}$ is able to approximate the target f ?

Generalization

How well the algorithm is able to pick a hypothesis from $\mathcal{H}$ that approximates well the target f ?

$$E_{out}(g) \leq E_{in}(g) + \Omega(N, \mathcal{H}, \delta)$$

Fitting / approximation $\longrightarrow$ the larger $\mathcal{H}$, the better the approximation

How well $\mathcal{H}$ is able to approximate the target f ?

$|E_{in}(g) - E_{out}(g)|$

Generalization $\longrightarrow$ the larger $\mathcal{H}$, the worse the generalization

How well the algorithm is able to pick a hypothesis from $\mathcal{H}$ that approximates well the target f ?

Bias-Variance: It is another model that has the approximation-generalization tradeoff structure

VC analysis: $E_{out}(g) \leq E_{in}(g) + \Omega$

E_{in} is computed with respect to a dataset D

Bias-variance analysis: $E_{out} = \textbf{bias} + \textbf{variance}$

bias refers to an average hypothesis $\overline{g}$

Bias-variance analysis for the regression problem

with mean squared error

$E_{out} = \textbf{bias} + \textbf{variance}$

$D = \{ (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)}) \}$

Start with E_{out}

$g^{(D)}$

$$\boxed{E_{\text{out}}(g^{(\mathcal{D})}) = \mathbb{E}_{\mathbf{x}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]}$$

$$\begin{aligned} \mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right] &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{x}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right] \\ &= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right] \end{aligned}$$

Now, let us focus on:

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$$

The average hypothesis

To evaluate $\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$

we define the 'average' hypothesis $\bar{g}(\mathbf{x})$:

$$\bar{g}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\mathbf{x})\right]$$

Imagine **many** data sets $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_K$

$$\bar{g}(\mathbf{x}) \approx \frac{1}{K}\sum_{k=1}^{K} g^{(\mathcal{D}_k)}(\mathbf{x})$$

Using $\bar{g}(\mathbf{x})$

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] = \mathbb{E}_{\mathcal{D}}\left[\left(\boxed{g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})} + \boxed{\bar{g}(\mathbf{x}) - f(\mathbf{x})}\right)^2\right]$$

$$= \mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2 + \left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\right.$$

$$\left. + \boxed{2\ \left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)\ \left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)}\ \right] \rightarrow 0$$

$$E_D\left[2\ \boxed{}\ \boxed{}\ \right]$$

$$2\ \boxed{}\ \underbrace{E_D\left[\ g^{(D)}(x) - \bar{g}(x)\ \right]}_{} = \mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right] + \left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2$$

Bias and variance

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x})-f(\mathbf{x})\right)^2\right]=\boxed{\underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x})-\bar{g}(\mathbf{x})\right)^2\right]}_{\text{var}(\mathbf{x})}+\underbrace{\left(\bar{g}(\mathbf{x})-f(\mathbf{x})\right)^2}_{\text{bias}(\mathbf{x})}}$$

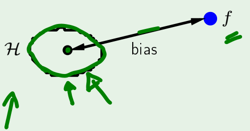
Therefore, $\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right]=\mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x})-f(\mathbf{x})\right)^2\right]\right]$

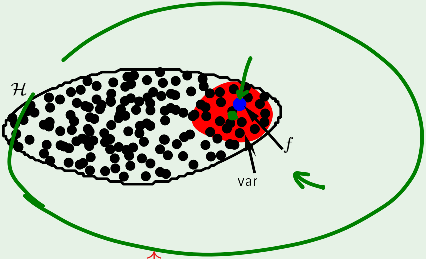
$$=\mathbb{E}_{\mathbf{x}}[\text{bias}(\mathbf{x})+\text{var}(\mathbf{x})]$$

$$=\quad\text{bias}\quad+\quad\text{var}$$

The tradeoff

$$\text{bias} = \mathbb{E}_{\mathbf{x}}\left[\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] \qquad \text{var} = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]\right]$$





$\downarrow \qquad \mathcal{H}\uparrow \qquad \uparrow$

VC analysis: $\boxed{E_{out}(g) \leq E_{in}(g) + \Omega}$

E_{in} is computed with respect to a dataset D

Bias-variance analysis: $E_{out} = \textbf{bias} + \textbf{variance}$

bias refers to an average hypothesis $\overline{g}$

(with respect to all datasets D of fixed size)

Example: sine target

f

$f : [-1, 1] \to \mathbb{R}$ $\quad f(x) = \sin(\pi x)$

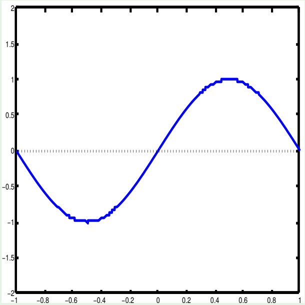
Only two training examples! $\quad N = 2$

Two models used for learning:

$\mathcal{H}_0$: $\quad h(x) = b$

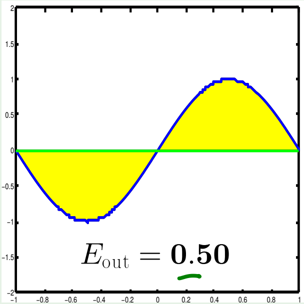
$\mathcal{H}_1$: $\quad h(x) = ax + b$

Which is better, $\mathcal{H}_0$ or $\mathcal{H}_1$?

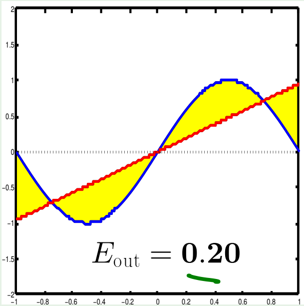


Approximation - $\mathcal{H}_0$ versus $\mathcal{H}_1$

$\mathcal{H}_0$

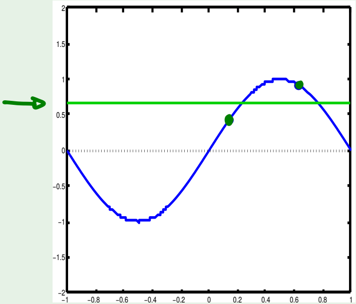


$\mathcal{H}_1$

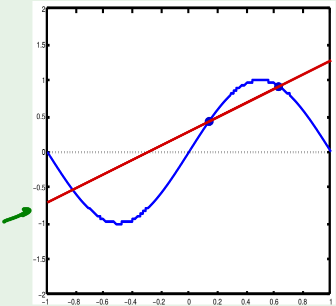


Learning - $\mathcal{H}_0$ versus $\mathcal{H}_1$

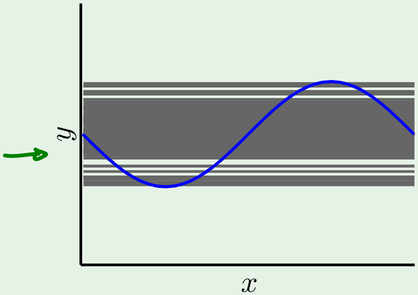
$\mathcal{H}_0$

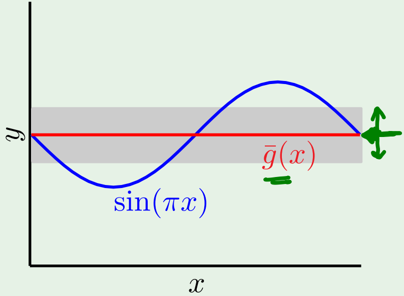


$\mathcal{H}_1$



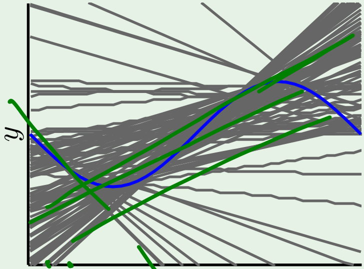
Bias and variance - $\mathcal{H}_0$

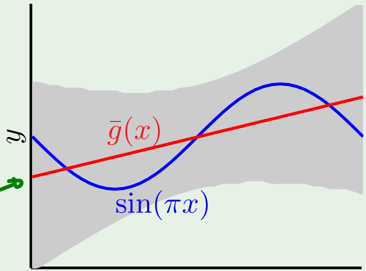




$ax+b$

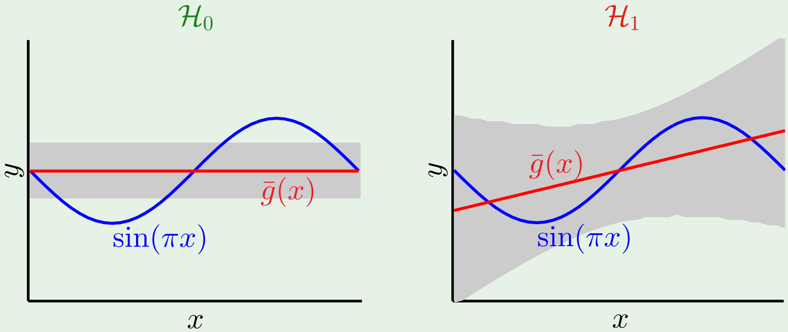
Bias and variance - $\mathcal{H}_1$





$N = 2$

and the winner is . . .



bias $= \mathbf{0.50}$ + var $= \mathbf{0.25}$ < bias $= \mathbf{0.21}$ + var $= \mathbf{1.69}$

© Creator: Yaser Abu-Mostafa - LFD Lecture 8

E_{out} = bias + variance

Learning curves

Expected E_{out} and E_{in}

Data set $\mathcal{D}$ of size N

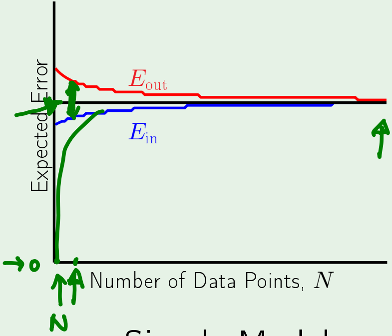
Expected out-of-sample error $\quad \mathbb{E}_{\mathcal{D}}[E_{\text{out}}(g^{(\mathcal{D})})]$

Expected in-sample error $\quad \mathbb{E}_{\mathcal{D}}[E_{\text{in}}(g^{(\mathcal{D})})]$

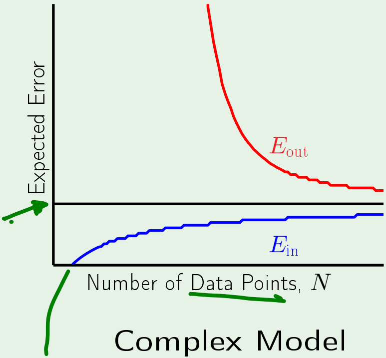
How do they vary with N?

$P(|E_{in} - E_{out}| > \varepsilon) \leq \text{bound}$

The curves



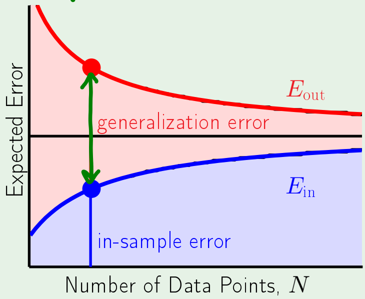
Simple Model



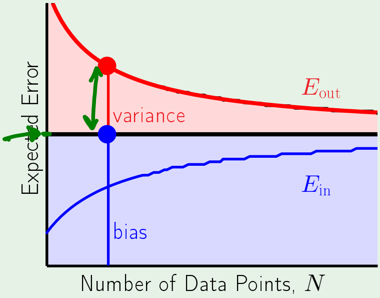
Complex Model

VC versus bias-variance

$|E_{in} - E_{out}|$



VC analysis



bias-variance

- VC and Bias-variance analysis decompose E_{out} in two terms

- They can be interpreted according to the **Approximation-generalization** tradeoff

$$
\left\{
\begin{array}{lcccl}
 & & \overset{\text{approx.}}{\downarrow} & & \overset{\text{generalization}}{\downarrow} \\
E_{out} & \leq & E_{in} & + & \Omega \quad \longleftarrow \text{VC} \\
E_{out} & = & \textbf{bias} & + & \textbf{variance}
\end{array}
\right.
$$

- Expressiveness of $\mathcal{H}$ should be matched to the amount of available data

$$E_{out} \approx E_{in} + \Omega.$$

Additional references

PAC model

- Valiant, Leslie (1984). "A theory of the learnable". Communications of the ACM. 27 (11): 1134–1142
- Tom Mitchell (1997). "Machine Learning", Chapter 7: Computational Learning Theory
- Vladimir Naumovich Vapnik (1995), The Nature of Statistical Learning Theory
- Vladimir Naumovich Vapnik (1998), Statistical Learning Theory
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2009), "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Second Edition

$$\rightarrow P(\underbrace{|E_{in} - E_{out}| < \varepsilon}) \geq \underbrace{1-\delta}$$